Advance Access publication August 3, 2009

Political Analysis (2009) 17:400–417
doi:10.1093/pan/mpp018

Testing the Accuracy of Regression Discontinuity Analysis Using Experimental Benchmarks

Donald P. Green
Department of Political Science, Institution for Social and Policy Studies, Yale University, 77 Prospect St., New Haven, CT 06511

Terence Y. Leong
Analyst Institute, 815 Sixteenth Street NW, Washington, DC 20006

Holger L. Kern
Institution for Social and Policy Studies, Yale University, 77 Prospect St., New Haven, CT 06511
e-mail: holger.kern@yale.edu (corresponding author)

Alan S. Gerber
Department of Political Science, Institution for Social and Policy Studies, Yale University, 77 Prospect St., New Haven, CT 06511

Christopher W. Larimer
Department of Political Science, University of Northern Iowa, 332 Sabin Hall, Cedar Falls, IA 50614

Regression discontinuity (RD) designs enable researchers to estimate causal effects using observational data. These causal effects are identified at the point of discontinuity that distinguishes those observations that do or do not receive the treatment. One challenge in applying RD in practice is that data may be sparse in the immediate vicinity of the discontinuity. Expanding the analysis to observations outside this immediate vicinity may improve the statistical precision with which treatment effects are estimated, but including more distant observations also increases the risk of bias. Model specification is another source of uncertainty; as the bandwidth around the cutoff point expands, linear approximations may break down, requiring more flexible functional forms. Using data from a large randomized experiment conducted by Gerber, Green, and Larimer (2008), this study attempts to recover an experimental benchmark using RD and assesses the uncertainty introduced by various aspects of model and bandwidth selection. More generally, we demonstrate how experimental benchmarks can be used to gauge and improve the reliability of RD analyses.

Authors' note: The authors are grateful to Mark Grebner, who designed and implemented the mailing campaign analyzed here, and Joshua Haselkorn, Jonnah Hollander, and Celia Paris, who provided research assistance.

Regression discontinuity (RD) is a research method that attempts to estimate the causal effect of an intervention by examining comparable observations in which the treatment is or is not administered (Thistlethwaite and Campbell 1960). When assignment to a treatment is determined by the value of a continuous observed variable, identification is achieved by comparing observations that lie on either side of a cutoff point that separates the treated from the untreated (Imbens and Lemieux 2008: 616). Recent years have seen a dramatic increase in the number and range of RD studies in social science. For example, Lee (2008) demonstrated the electoral advantage enjoyed by incumbents using an RD analysis that compared the vote share of the party that narrowly won the preceding election to the vote share of the party that narrowly lost (for other electoral applications, see Butler and Butler 2006; Hainmueller and Kern 2008). Pettersson-Lidbom (2004) estimated the budgetary effects of increasing the number of municipal representatives by comparing towns that acquire additional representatives based on a population-based formula. In these applications, a rigidly applied rule divides a continuous variable—vote share, population—allowing researchers to compare outcomes among what are otherwise similar observations on either side of the cutoff.

The attraction of RD analysis is its capacity to extract unbiased estimates of causal effects from nonexperimental data. In domains where random assignment of treatments is impractical, RD offers a way to approximate random assignment in the vicinity of the discontinuity. As Cook and Wong (2009) stress, not all discontinuities are suitable for this kind of analysis; in some cases, actors may be aware of a discontinuity (e.g., a change in income tax brackets) and adjust their behavior accordingly. In many applications, the targeting criteria involve variables whose distributions are too coarse to allow for a convincing comparison of treated and untreated observations. Nevertheless, it appears that the number of potentially fruitful applications of RD is quite large, and social scientists have only begun to tap the supply of applications in which decision makers partition a continuous variable and administer a treatment to those on one side of the partition.

However, even in instances where RD designs seem apt, there remains the question of how best to analyze data associated with a RD. Researchers have considerable discretion when it comes to two aspects of data analysis: the size of the window surrounding the point of discontinuity and the statistical model used to estimate the treatment effect. The narrower the window, the less chance that omitted variables will bias the estimated treatment effect. On the other hand, the narrower the window, the smaller the number of available observations. In the limiting case, the exact point of discontinuity is a set of measure zero: inferences will be unbiased but, without data, the analyst confronts infinite sampling variability. Widening the window expands the number of observations, but doing so also introduces the risk of bias due to omitted determinants of the outcome that are correlated with the treatment. In an attempt to reduce the threat of bias, researchers often control for the variables along which the discontinuity occurs, but the adequacy of this approach depends on the particular application. Taken together, these two specification issues present the researcher with a tradeoff: use a narrow window and put up with sampling variability (and potentially uninformative results) or expand the window and risk introducing bias.

How to resolve this conundrum will doubtless depend on the application at hand, but researchers may also benefit from studies that compare estimates generated by RD to results from an experimental benchmark. Although this exercise, pioneered by LaLonde (1986), has become widespread as a means of analyzing the adequacy of a variety of statistical methods used to analyze observational data, it has rarely been used to assess and guide RD analysis (Buddelmeyer and Skoufias 2003; Black, Galdo, and Smith 2005). The one exception of which we are aware in political science is Nickerson (2007), which compares an experimental evaluation of an age-targeted voter mobilization campaign to a RD

analysis based on a sharp age cutoff. Our paper builds on this work using a large data set that enables us to simulate discontinuities and evaluate alternative estimation approaches.

In the application presented here, a campaign consultant constructed a target list of voters for a direct mail campaign.[1] The target voters were randomly divided into control and treatment groups, enabling Gerber, Green, and Larimer (2008) to estimate experimentally the effects of four different types of mailings on voter turnout. The question for the current paper is whether one can recover the estimated treatment effects from this experiment by means of a RD analysis using a simulated discontinuity. Like many applications of RD, ours presents the problem that data in the immediate vicinity of the discontinuity are sparse, and so the analyst must decide how much to widen the window around the discontinuity and how to model the comparison between groups that do or do not receive the treatment. After providing a brief overview of RD analysis, we estimate an experimental benchmark and attempt to recover it using various RD estimation approaches. Although some RD estimation procedures perform well, even the most promising approaches sometimes produce misleading estimates and SEs. We conclude by discussing the practical implications of uncertainty about how best to specify RD models and by suggesting avenues for further empirical investigation using experimental benchmarks.

1 Causal Inference from Sharp RD Designs

RD designs are typically explicated in terms of a potential outcomes framework that defines a treatment effect as the difference between two quantities: the outcome when observation i is exposed to the treatment, $Y_i(1)$, and the outcome when the observation is not exposed to the treatment, $Y_i(0)$ (Hahn, Todd, and Van der Klaauw 2001).

Because we cannot simultaneously observe voters in their treated and untreated states, we must rely on average responses of observations that are randomly or near-randomly assigned to treatment and control conditions. Let T_i denote the treatment received by each observation, with $T_i = 1$ if the unit received the treatment and $T_i = 0$ otherwise. The observed outcomes are therefore given by:

$$Y_i = (1-T_i)Y_i(0)+T_iY_i(1) = \begin{cases} Y_i(0) & \text{if } T_i = 0 \\ Y_i(1) & \text{if } T_i = 1 \end{cases}. \quad (1)$$

In other words, we observe treated outcomes for observations receiving the treatment and untreated outcomes for observations not receiving the treatment.

The challenge then is to construct a comparison between treated and untreated units that recovers the treatment effect without being confounded by preexisting differences that are correlated with receipt of the treatment. One way is through random assignment of the treatment (Rubin 1974). RD provides another way. Let c represent a cutoff point such that $T_i = 1$ when some covariate X is greater than or equal to c and $T_i = 0$ when X is less than c. This situation is known as a ''sharp'' RD because T_i is a deterministic function of X, the so-called forcing variable. In this case, the average treatment effect at $X = c$ is

$$\tau = E[Y(1)-Y(0)|X = c] = E[Y(1)|X = c]-E[Y(0)|X = c].$$

Imbens and Lemieux (2008: 618) show that under fairly weak continuity assumptions, the average treatment effect in the vicinity of the cutoff point can be estimated by

[1]Using the criteria discussed below, 344,081 target voters were selected from a list of 6,710,669 registered voters from the state of Michigan. Here, we focus on a subset of the target sample, single-voter households who were assigned to either the control group or one of the experimental treatment groups.

comparing average outcomes on either side of the discontinuity. Continuity, although a weak requirement in principle, may be an invalid assumption in practice. It may be that decision makers are aware of the discontinuity threshold and manipulate outcomes to offset its influence. As noted earlier, discontinuous changes in marginal tax rates may not be suitable for analysis because taxpayers may behave strategically to fall just short of the cutoff point. Few election campaign interventions suffer from this problem, as campaign targeting is typically done using eligibility criteria of which voters are unaware.

2 Estimating an Experimental Benchmark: An Example Involving Direct Mail and Voter Mobilization

The benchmark used here comes from a study conducted by Gerber, Green, and Larimer (2008), which gauged the effects of various direct mailings on voter turnout among Michigan residents prior to the August 2006 primary election. This section provides a brief overview of that experiment.

2.1 *Setting*

The August 2006 primary was a statewide election with a wide range of offices and proposals on the ballot, most of which were limited to counties, cities, and local districts. There were no important contested statewide Democratic primary elections and just one competitive primary contest for U.S. Senate and U.S. House, both on the Republican side. Michigan voters are allowed to vote in either the Democratic primary or the Republican primary, but not both. The voter's choice of party is secret under Michigan law, and there is no party registration. For those intending to vote as Democrats, there was little reason to vote in the 2006 primary apart from the occasional nonpartisan judicial race or contested local office. Voter turnout in the August 2006 primary was 1,282,203, or 17.7% of registered voters.

2.2 *Treatments*

Each household in the treatment group received one of four mailings. The appendix to Gerber, Green, and Larimer (2008) shows examples of each type. Priming voters to think about their civic duty is common to all the treatment mailings. All four treatments carry the message "DO YOUR CIVIC DUTY—VOTE!" The first type of mailing ("Civic Duty") provides a baseline for comparison with the other treatments because it does little besides emphasize civic duty. Households receiving this type of mailing were told, "Remember your rights and responsibilities as a citizen. Remember to vote." The second mailing adds to this civic duty baseline a mild form of social pressure, in this case, observation by researchers. Households receiving the "Hawthorne effect" mailing were told "YOU ARE BEING STUDIED!" and informed that their voting behavior would be examined by means of public records. The degree of social pressure in this mailing was, by design, limited by the promise that the researchers would neither contact the subject nor disclose whether the subject voted. The "Self" mailing exerts more social pressure by informing recipients that who votes is public information and listing the recent voting record of each registered voter in the household. The word "Voted" appears by names of registered voters in the household who actually voted in the 2004 primary election and the 2004 general election, and a blank space appears if they did not vote. The mailing informed voters that after the primary election, "we intend to mail an updated chart" filling in whether the recipient voted in the August 2006 primary. The Self condition thus combines the external monitoring of the Hawthorne condition with actual disclosure of voting records. The fourth mailing, "Neighbors," exerts still more social pressure by listing not only the household's voting records but also the

voting records of those living nearby. Like the Self mailing, the Neighbors mailing informed the recipient that ''we intend to mail an updated chart'' after the primary, showing whether members of the household voted in the primary and who among their neighbors had actually voted in the primary. The implication is that members of the household would know their neighbors' voting records and their neighbors would know theirs.

As Gerber, Green, and Larimer (2008) demonstrate, the treatment effects grow larger as social pressure increases. The most dramatic effect is associated with the Neighbors treatment. Although less dramatic effects are also susceptible to RD analysis, the advantage of working with a powerful experimental intervention is that it lends itself to more instructive graphical presentation. For this reason, the analysis below focuses on the comparison between the control group and the Neighbors group.

2.3 *Selection of the Treatment Group*

The targeting criteria used in this mailing campaign were developed by a political consultant, Mark Grebner, a longtime veteran of political campaigns involving direct mail. Like many political consultants, he uses targeting criteria based on a combination of address information readily available from the Qualified Voter File (QVF) and a set of proprietary indices of partisanship and voting behavior developed by his consulting firm. Grebner's targeting objective was to direct mailings to those who were thought to be especially responsive to them. Mailings were therefore sent to voters whose expected probability of voting was deemed to be moderate. Those believed to be strong Democrats were excluded on the grounds that they had little chance of voting in an election that was meaningful mainly to Republicans. Absentee voters were excluded because they were believed to vote early, before the receipt of these mailings. Sparsely populated streets were excluded because the Neighbors treatment requires the voting histories of several neighbors. Apartment addresses were excluded because apartment numbers are sometimes unreliable, and it is hard to be certain which voters belong to the same household.

The criteria used to restrict the target sample were therefore as follows.[2] Voters were targeted if (1) their forecasted probability of voting was between 30% and 80%, (2) their forecasted probability of voting for a Democrat in a general election was less than 59%, (3) their forecasted probability of voting by absentee ballot was less than 60%, (4) they lived on a street segment in which there were at least 10 voter households, and (5) the ratio of apartment addresses to voter households on a given street segment was less than 0.09. If a voter failed to meet any of these criteria, he or she was excluded from the target sample. For purposes of this analysis, we restrict our attention to one-voter households because doing so allows us to sidestep complications that are irrelevant to the assessment of RD analysis.[3] Note that Grebner's targeting criteria are broadly illustrative of the kind of eligibility rules that are used in campaigns. They lend themselves to a RD analysis, albeit an unusually complex one given the number and intricacy of the selection criteria (see

[2]This summary corrects some minor errors in the description given in Gerber, Green, and Larimer (2008). Note that Grebner also limited the targeted mail to those who had voted in the November 2004 election on the grounds that, given the very high turnout among registered voters in that election, anyone failing to do so was probably no longer living at the address listed in the QVF.

[3]The two complications are as follows. First, all individuals at a given address were assigned as a cluster to one of the treatment groups. Including multivoter households would require correcting the SEs for clustered random assignment. Second, Grebner included individuals in the target sample if they either met all the selection criteria or at least one of their housemates did. Thus, some multivoter households contain individuals on opposite sides of the cutoff thresholds. Neither complication arises if we restrict attention to single-voter households in the QVF.

Green et al. 2008). Here, we use the experiment in a different manner, simulating much simpler targeting criteria that classify voters along a single dimension.

3 Benchmark Estimates

Using voter turnout data from public records, we can estimate the average treatment effect of the Neighbors mailing for the entire sample. The control group in our study (N = 24,964) voted at a rate of 32.64%. By comparison, turnout among those assigned to the Neighbors mailing (N = 5074) is 42.29%, which implies a 9.65 percentage-point treatment effect. Due to the large sample sizes in each experimental group, these treatment effects may be estimated with a high degree of precision using a linear probability model, with a robust SE of 0.75 percentage points. Note that this is the average treatment effect for the entire sample, not the treatment effect at a particular discontinuity.

There are many ways to simulate a discontinuity using these data; following Nickerson (2007), we consider a hypothetical targeting rule based on age. For example, the experimental data used here could be divided according to an age threshold such that those younger than 55 years receive a mailing; those 55 years and older do not.[4] This faux discontinuity is readily simulated in this case by restricting the data set to 2642 observations in the treatment group who are younger than 55 years and 12,075 observations in the control group who are 55 years and older.

Unlike an actual mail campaign, our experiment gives us an empirical glimpse at the counterfactuals: Sampling error aside, we see how those 55 years and older would have behaved if they had received mail, and we see how those younger than 55 years would have behaved if they had not received mail. In essence, our experimental mail campaign contains everything that we would observe as part of an actual campaign plus the counterfactuals.

Figure 1 shows in red the observed data on each side of the discontinuity. The data have been aggregated by year of age, and the group average is plotted using circles whose size is proportional to the number of observations. The blue circles depict the unobserved counterfactual outcomes. To the left of the age cutoff, the blue circles indicate how those eligible for mail would have behaved had they not received it; to the right of the cutoff, the blue circles show how those ineligible for mail would have behaved had they received it. The analyst of an actual RD would only observe the data depicted in red.

The benchmark regression model attempts to estimate the effect of the Neighbors treatment at the point of the discontinuity. To do so, we add Age as a covariate and subtract 55 from it so that Age = 0 at the discontinuity. Age is interacted with the Neighbors treatment so that the "main effect" of Neighbors (β_1) represents the effect of this mailing when Age = 0 (i.e., at the point of discontinuity). Anticipating the flexible polynomials used in the RD models below, we also include higher powers of Age and interactions between them and Neighbors. The regression model with a fourth-order polynomial in Age that allows for different curves on each side of the discontinuity, for example, looks as follows:

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_1 \text{Neighbors}_i + \beta_2 \text{Age}_i + \beta_3 \text{Age}_i \text{Neighbors}_i + \beta_4 \text{Age}_i^2 \\ & + \beta_5 \text{Age}_i^2 \text{Neighbors}_i + \beta_6 \text{Age}_i^3 + \beta_7 \text{Age}_i^3 \text{Neighbors}_i \\ & + \beta_8 \text{Age}_i^4 + \beta_9 \text{Age}_i^4 \text{Neighbors}_i + u_i. \end{aligned} \tag{2}$$

[4]This particular age cutoff is arbitrary, and one could use these data to investigate a range of hypothetical cutoffs as discussed below. In the statistical analyses that follow, age is measured in days, and the cutoff is based on a voter's age on Election Day. We exclude observations for which the QVF does not list a full birth date.

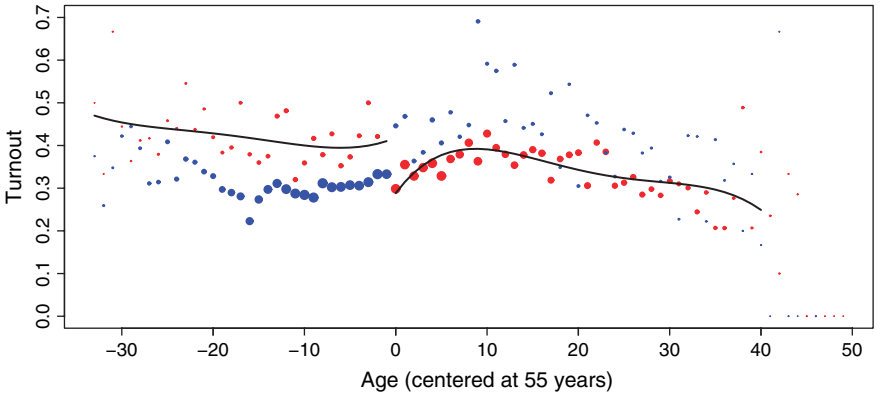

Fig. 1 Illustration of RD using age as a forcing variable. The red circles depict average voting rates among observed voters, grouped by year of age, which has been rescaled so that zero (age 55) is the point of discontinuity. The blue circles depict average voting rates among counterfactual voters. The red circles to the left of the age cutoff (where age equals 0) represent the treatment group, which received the experimental mailings. The red circles to the right of the cutoff represent the control group, which received no experimental mailings. The size of the circles is proportional to the number of observations in each age group.

The fitted lines in Fig. 1 illustrate how this model works. The vertical distance at the point of discontinuity is the estimate of β_1. Other model specifications with fewer polynomials are subsets of equation (2). For example, a linear specification simply constrains $\beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$.

Column 1d of Table 1 shows that the benchmark is 10.0 percentage points with a robust SE of 1.1 percentage points. Note that this estimated local effect is quite similar to the estimated global effect of 9.7 with a robust SE of 0.8 reported in column 1a. Specifications with different order polynomials produce similar results. For example, including up to third order polynomials and their interactions produces a benchmark estimate of 10.3 with a robust SE of 1.0. Choosing among these estimates has no effect on the conclusions below, and so we will use the estimates generated by the fourth-order specification depicted in equation (2). We now assess the degree to which RD estimates recover this experimental benchmark. Bearing in mind the fact that RD analysts will not ordinarily have knowledge of an experimental benchmark to guide their modeling, we pay special attention to the kinds of model and bandwidth selection criteria that researchers may use when choosing among alternative RD specifications.

4 Recovering the Benchmark: RD Estimation Using Polynomials

Our first attempt to recover the experimental benchmark uses the entire sample (the maximum bandwidth) and varies the specification of the polynomial regression model. The RD model treats Age as the forcing variable. Again, Age is centered at 55. An RD model with a fourth-order polynomial in Age that allows for different curves on each side of the discontinuity is the same as equation (2). The parameter of interest is β_1, which represents the effect of the Neighbors treatment when Age $= 0$ (i.e., at the point of discontinuity). The difference between the benchmark regression and the RD analysis is the set of observations on which the model is estimated: The benchmark analysis uses all the data, whereas the RD analysis is restricted to treatment observations to the left of the cutoff and control observations to the right of the cutoff.

Table 1 Comparison between RD estimates and experimental benchmarks, full sample

	Full Sample							
	Benchmark				*RD*			
	(1a)	*(1b)*	*(1c)*	*(1d)*	*(2a)*	*(2b)*	*(2c)*	*(2d)*
Neighbors	9.65**	10.43**	10.26**	10.0**	1.77	7.36**	10.42**	10.33**
treatment	(0.75)	(0.97)	(0.99)	(1.11)	(1.89)	(2.70)	(3.51)	(4.28)
(robust SE)								
Age		0.13**	0.039**	0.45**	−0.17**	0.66**	1.26**	1.79**
		(0.023)	(0.036)	(0.0026)	(0.042)	(0.14)	(0.31)	(0.55)
Age^2		−0.0054**	0.0016	−0.0058**		−0.025**	−0.074**	−0.12**
		(0.0012)	(0.0014)	(0.0026)		(0.0040)	(0.019)	(0.055)
Age^3			−0.00054**	−0.00072**			0.00088**	0.0029
			(0.000059)	(0.000083)			(0.00033)	(0.0020)
Age^4				0.0000097**				−0.000026
				(0.0000028)				(0.000024)
Age × Neighbors		−0.051	−0.0055	−0.032	0.032	−0.48	−0.71	−1.75
		(0.057)	(0.09)	(0.11)	(0.15)	(0.47)	(1.06)	(1.99)
Age^2 × Neighbors		−0.0031	−0.0014	0.0016		0.038**	0.13	0.083
		(0.0028)	(0.0035)	(0.0065)		(0.017)	(0.088)	(0.277)
Age^3 × Neighbors			−0.000084	−0.000016			0.00016	−0.0072
			(0.00015)	(0.00021)			(0.0020)	(0.014)
Age^4 × Neighbors				−0.0000040				−0.000066
				(0.0000071)				(0.00024)
Observations	30,038	30,038	30,038	30,038	14,717	14,717	14,717	14,717
Squared error					67.73	6.97	0.18	0.11
MSE					71.31	14.26	12.50	18.43

Note. Dependent variable is voter turnout. Age has been centered so that it is zero at the point of discontinuity (55 years). Table entries are least squares regression estimates and robust SEs. Squared error is the squared difference between the estimates in columns 2a–d and the benchmark estimate of 10.0. MSE is mean squared error, defined as squared error plus the square of the SE.

*$p < .10$, **$p < .05$ (two-tailed test).

Results are reported in Table 1, columns 2a–d. As intuition suggests, the SE associated with the estimated treatment effect increases as additional polynomial terms are added to the specification. For this reason, the researcher who is unaware of the experimental benchmark may be attracted to the linear specification in column 2a. However, the linear specification produces an estimate ($b = 1.8$, robust SE = 1.9) that is more than four SEs away from the benchmark. The accuracy of the RD estimator improves as a quadratic term is added ($b = 7.4$, SE = 2.7). The squared difference between the estimate and the benchmark declines from 67.7 to 7.0 as a quadratic term and interaction are added, suggesting that the relationship between age and vote is nonlinear on either side of the cutoff. Accuracy continues to improve as higher order terms are introduced, with the cubic ($b = 10.4$, SE = 3.5) and quartic ($b = 10.3$, SE = 4.3) RD specifications recovering the benchmark estimates almost exactly. The downside of the cubic and quartic specification is the sharp increase in SEs. By the mean squared error criterion (bias squared plus variance), the cubic specification offers the best balance of bias and precision.

The choice among specifications may contribute to the overall level of uncertainty associated with RD estimates, and so it is instructive to consider the sampling distribution that results from specification searches. One commonly used rule of thumb is to start with a high-order polynomial and successively pare down the specification depending on whether the highest order terms (and associated interactions) are statistically significant at the $p < .05$ level. The search stops when the highest order term or interaction is significant. In Table 1, this procedure would settle on the cubic specification, as none of the quartic terms is significant. It turns out that there is little sampling variability in this aspect of the specification search: applying this model selection procedure to 1000 bootstrap samples generates estimates with a SD of 3.9 percentage points, as opposed to the nominal SE of 3.5 from the cubic regression.

In addition to choosing among alternative polynomial specifications, researchers must decide how narrowly to restrict the sample on either side of the cutoff point. Narrowing the window around the cutoff lowers the risk of bias but reduces the number of observations. Narrowing the window presumably makes the regression more plausibly linear in the vicinity of the discontinuity, which reduces the need for higher order polynomials and attendant problems of collinearity. Assessing the robustness of results is an important aspect of RD analysis, as Imbens and Lemieux (2008: 633) warn:

> Irrespective of the manner in which the bandwidth is chosen, one should always investigate the sensitivity of the inferences to this choice, for example, by including results for bandwidths twice (or four times) and half (or a quarter of) the size of the originally chosen bandwidth. Obviously, such bandwidth choices affect both estimates and standard errors, but if the results are critically dependent on a particular bandwidth choice, they are clearly less credible than if they are robust to such variation in bandwidths.

To illustrate the consequences of reducing bandwidth (defined as the distance between the cutoff and the lowest or highest age that is included in the analysis), we consider three illustrative windows in Table 2: ages falling within 5 years of the cutoff, within 10 years of the cutoff, and within 20 years of the cutoff. The experimental benchmark, the estimated treatment effect at age 55, remains the same; the question is how well the RD models perform as the window size changes. Restricting the sample to ages 50–60 causes the N to drop by about 75%, and the estimates have larger SEs. The narrow window does not appear to lead to especially accurate estimates. More troubling perhaps is the fact that the statistical significance of the higher order terms does not provide reliable guidance as to the most accurate specification. For example, third-order terms are significant, yet the estimated effect of the mailing based on this RD specification is 19.4 with a SE of 8.4.

Table 2 RD estimates, restricted samples

	Restricted samples											
	5 years				*10 years*				*20 years*			
	(1a)	*(1b)*	*(1c)*	*(1d)*	*(2a)*	*(2b)*	*(2c)*	*(2d)*	*(3a)*	*(3b)*	*(3c)*	*(3d)*
Neighbors	7.64*	3.90	19.41**	29.58**	9.92**	7.21	8.10	5.52	5.63**	9.79**	10.52**	6.60
treatment	(4.18)	(6.38)	(8.40)	(10.57)	(2.98)	(4.44)	(5.97)	(7.53)	(2.18)	(3.24)	(4.28)	(5.33)
(robust SE)												
Age	0.53	4.66**	6.09	20.68*	0.82**	0.78	2.45	6.50	0.26**	1.26**	1.71**	0.17
	(0.56)	(2.27)	(5.74)	(11.52)	(0.23)	(0.87)	(2.14)	(4.23)	(0.090)	(0.034)	(0.81)	(1.57)
Age^2		−0.81*	−1.52	−14.58		0.0046	−0.43	−2.28		−0.053**	−0.11	0.25
		(0.44)	(2.67)	(9.38)		(0.086)	(0.52)	(0.05)		(0.018)	(0.10)	(0.34)
Age^3			0.094	4.13			0.029	0.322			0.0021	−0.027
			(0.35)	(2.81)			(0.035)	(0.27)			(0.0034)	(0.026)
Age^4				−0.40				−0.014				0.00076
				(0.28)				(0.014)				(0.00068)
Age ×	−1.0	−13.74**	20.89	32.24	−0.46	−1.97	−4.17	−17.30*	−0.29	−0.93	−1.37	−2.36
Neighbors	(1.43)	(5.85)	(14.37)	(28.96)	(0.51)	(2.02)	(5.20)	(10.35)	(0.21)	(0.78)	(1.91)	(3.80)
Age^2 ×		−0.89	17.87**	54.12**		−0.16	0.14	−2.05		0.072*	0.13	−0.82
Neighbors		(1.11)	(6.63)	(23.36)		(0.20)	(1.21)	(4.16)		(0.40)	(0.23)	(0.80)
Age^3 ×			2.31**	5.45			−0.038	−0.96			−0.0020	−0.020
Neighbors			(0.86)	(6.96)			(0.079)	(0.62)			(0.0079)	(0.062)
Age^4 ×				1.12				−0.016				−0.0020
Neighbors				(0.68)				(0.031)				(0.0016)
Observations	3813	3813	3813	3813	6905	6905	6905	6905	11,243	11,243	11,243	11,243
Squared error	5.57	37.21	88.55	383.38	0.0064	7.78	3.61	20.07	19.10	0.044	0.27	11.56
MSE	23.04	77.91	159.11	495.10	8.89	27.50	39.25	76.77	23.85	10.54	18.59	39.97

Note. Dependent variable is voter turnout. Age has been centered so that it is zero at the point of discontinuity (55 years). Table entries are least squares regression estimates and robust SEs. Squared error is the squared difference between the estimates in columns 2a–d and the benchmark estimate of 10.0. MSE is mean squared error, defined as squared error plus the square of the SE.

*$p < .10$, **$p < .05$ (two-tailed test).

Widening the window to ages 45–65 leads to improved estimates and smaller SEs. Using roughly half of the available data, the linear specification produces an estimate of 9.9, which is very close to the experimental benchmark, with a SE of 3.0. None of the higher order terms is significant, which means that a conventional specification search would produce an accurate result in this instance. Widening the window still further to admit ages 35–75 also produces adequate results. The model with quadratic terms produces accurate estimates ($b = 9.8$, SE $= 3.2$) and is the one selected by a specification search guided by the significance of higher order terms.

These results call attention to some of the tradeoffs that confront the RD analyst: widening the window around the cutoff point increases sample size and admits observations that have the potential to assist in the estimation of curves on either side of the discontinuity. On the other hand, the higher order polynomials that make these curves flexible also increase collinearity among the right-hand-side variables and thus increase the SEs surrounding the estimated treatment effect. Narrowing the window around the discontinuity facilitates a locally linear regression, but the precision of this regression depends on whether sufficient observations lie near the discontinuity. The next section considers efforts to automate the selection of the window around the cutoff.

5 RD Estimation Using Local Regression

An alternative to polynomial regression is to fit local linear regression models (Loader 1999) in the vicinity of the cutoff. The specification is a simplification of the model in equation (2), this time excluding higher order polynomials and their interactions:

$$Y_i = \beta_0 + \beta_1 \text{Neighbors}_i + \beta_2 \text{Age}_i + \beta_3 \text{Age}_i \text{Neighbors}_i + u_i. \quad (3)$$

Again, the parameter of interest is β_1, the effect of the treatment evaluated at the point of discontinuity (where Age $= 0$).

This modeling approach requires the analyst to make two important decisions concerning bandwidth and weights. In an effort to automate the process of bandwidth selection, researchers have developed two algorithms designed to balance the tradeoff between sampling variability and bias. Ludwig and Miller (2007) and Imbens and Lemieux (2008) have proposed a ''leave one out'' cross-validation procedure aimed specifically at estimating the regression function at the cutoff. To see how well a local linear regression with bandwidth h fits the data, they run a local linear regression for each observation i with i left out of the sample and then use the resulting coefficient estimates to predict the value of Y_i at X_i. Mimicking the fact that RD estimates are based on regression estimates at a boundary, the regressions are estimated using only observations to the left of i (for i below the cutoff) or the right of i (for i above the cutoff). Repeating this exercise N times produces a set of predicted values of Y_i that can be compared with the actual values of Y_i. The final ''cross-validated'' bandwidth is then picked by choosing the value of h that minimizes the mean square of the difference between the predicted and the actual values of Y_i.

An alternative approach uses asymptotic theory to derive the value of h. Imbens and Kalyanaraman (2009) derive the optimal bandwidth for the RD setting—optimal in the sense that it minimizes mean squared error—and propose an empirical procedure for bandwidth selection. The specific algorithm is complex, but the essential idea is to increase the size of h as the variance in outcomes at the cutoff increases, as the density of the forcing variable at the cutoff diminishes, and as the shape of the curves on opposite sides of the cutoff becomes increasingly symmetrical.

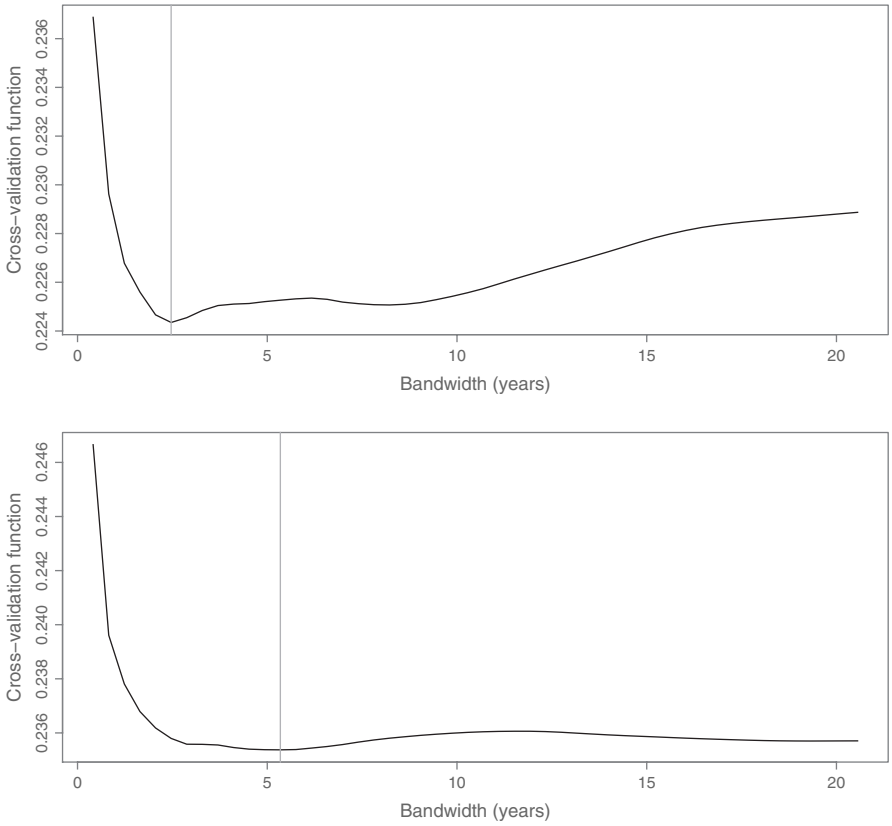

Fig. 2 Bandwidth selection using cross-validation, by sample size restriction. Cross-validation is used to select a bandwidth for local linear regressions. The upper graph shows the cross-validation function when the sample is restricted to the 10% of observations lying closest to the age cutoff. The vertical line denotes the minimum at $h = 2.47$ years. The lower graph shows the cross-validation function when the sample is restricted to the 50% of observations closest to the cutoff. The vertical line denotes the minimum at $h = 5.34$ years.

In our application, the optimal algorithm selects a bandwidth of 11.7 years. Bootstrapping the sample and recalculating the optimal bandwidth 10,000 times suggests that the value of h has a SE of 2.5 years, which is reassuringly small. The bandwidth calculation using cross-validation is less robust and involves another decision on the part of the researcher, namely, what proportion of the sample in the tails of the forcing variable to exclude prior to the search for h. Imbens and Lemieux (2008: 629) suggest discarding between 50% and 95% of the data before searching for the best-fitting value of h. When we exclude the 50% of the sample lying farthest from the age cutoff, the bandwidth is estimated to be 5.3 years (see Fig. 2). When we exclude 90% of the sample, the bandwidth falls to 2.5 years.

The second decision involves the use of weights. Intuitively, it makes sense to weight more heavily those observations falling closest to the cutoff. The relative size of the weights near and far from the threshold could follow any arbitrary monotonic function, but we follow Imbens and Kalyanaraman (2009) in using a triangular kernel for which weights decrease linearly from 1 (at the cutoff) to 0 (at h). The literature suggests that

the choice of kernel (e.g., rectangular, Gaussian, or Epanechnikov) makes little difference in practice (Lee and Lemieux 2009: 39).

Table 3 presents the results of local linear regressions using either the cross-validated or the asymptotically optimal bandwidth sizes h. The local linear regression using the asymptotically optimal bandwidth of 11.7 years on each side of the discontinuity generates an estimate of 9.4 percentage points with a robust SE of 3.0. As Fig. 3 illustrates, the local linear estimate using an optimal bandwidth comes very close to the benchmark because the selected bandwidth is nearly ideal. By contrast, the local linear regression using the cross-validation bandwidth of 2.5 years produces an estimate of 14.8 with a robust SE of 6.7; using a bandwidth of 5.3 generates an estimate of 6.4 with a robust SE of 4.5. In terms of mean squared error, the estimate using the optimal bandwidth is comparable or superior to the best polynomial regression models presented in Table 1, and the procedure used to generate these estimates is more clearly defined and leaves fewer decisions to the researcher. The two estimates generated using cross-validated bandwidths are relatively inaccurate, and the procedure that generates the bandwidths is less well defined.

6 Discussion

In principle, RD analysis offers social scientists the opportunity to draw sound causal inferences from nonexperimental data. Although RD methods were proposed decades ago, applications of this method remained relatively rare until the 1990s. Lee and Lemieux's (2009) recent review of RD applications charts the explosive growth of this method over the past decade. This growth is expected to continue in political science, where potential applications abound due to the rigid manner in which institutional rules allocate representation and government resources. The growth should be particularly vigorous in the study of campaign communications, where targeting criteria divide individuals and precincts in arbitrary ways (Gerber, Kessler, and Meredith 2009).

As RD analysis gains prominence, it becomes increasingly important to evaluate its performance through benchmarking exercises. Unlike simulated data, experimental data involve actual outcomes and realistic forcing variables. Although each application has its own idiosyncrasies, the process of recovering benchmarks from real data calls attention to the role of discretion in RD analysis. The analyst does not know ex ante which RD model specification is appropriate, and the results may be sensitive to choice of functional form. Bandwidth selection is another source of uncertainty. Indeed, one reason to prefer narrow bandwidths is that they tend to facilitate local linear regression. However, narrow bandwidths may generate estimates that are too uncertain to be useful, and one could imagine the dilemma that an analyst would face given the marked variation in results as bandwidths expand.

Our results illustrate some of these dilemmas.[5] Selecting the estimate with the smallest estimated SE was sometimes a poor guide to minimizing the mean squared error with which the benchmark was estimated. Specification searches based on statistical significance of the higher order terms in a polynomial specification performed adequately in some instances and not others. Maximizing the statistical significance of the estimate

[5]For the sake of brevity, the analysis presented above did not consider two other important sources of analyst discretion: whether to include covariates and whether to heed the results of placebo tests whereby outcomes that could not be affected by the RD intervention are used as dependent variables. Because our cutoff is by construction exogenous, we also dispense with diagnostic tests such as the examination of the density of Age in the vicinity of the discontinuity (McCrary 2008).

Table 3 Comparison between local linear regression estimates and experimental benchmarks, with different bandwidth selection algorithms

	Benchmark				*RD, local linear regression*		
	(1a)	*(1b)*	*(1c)*	*(1d)*	*Cross-validated bandwidth, based on 10% of sample (h = 2.5)*	*Cross-validated bandwidth, based on 50% of sample (h = 5.3)*	*Optimal bandwidth (h = 11.7)*
Neighbors treatment	9.65**	10.43**	10.26**	10.0**	14.75**	6.44	9.41**
(robust SE)	(0.75)	(0.97)	(0.99)	(1.11)	(6.65)	(4.47)	(3.03)
Age		0.13**	0.039**	0.45**	2.66	−0.73	0.096
		(0.023)	(0.036)	(0.0026)	(2.02)	(0.63)	(0.194)
Age^2		−0.0054**	0.0016	−0.0058**			
		(0.0012)	(0.0014)	(0.0026)			
Age^3			−0.00054**	−0.00072**			
			(0.000059)	(0.000083)			
Age^4				0.0000097**			
				(0.0000028)			
Age × Neighbors		−0.051	−0.0055	−0.032	−0.82	1.21*	0.252
		(0.057)	(0.09)	(0.11)	(2.20)	(0.69)	(0.213)
Age^2 × Neighbors		−0.0031	−0.0014	0.0016			
		(0.0028)	(0.0035)	(0.0065)			
Age^3 × Neighbors			−0.000084	−0.000016			
			(0.00015)	(0.00021)			
Age^4 × Neighbors				−0.0000040			
				(0.0000071)			
N (unweighted)	30,038	30,038	30,038	30,038	14,717	14,717	14,717
N with weights > 0	30,038	30,038	30,038	30,038	1844	4053	7796
Squared error					22.56	12.67	0.35
MSE					66.78	32.65	9.53

Dependent variable is voter turnout. Age has been centered so that it is zero at the point of discontinuity (55 years). Table entries are weighted least squares regression estimates and robust SEs. Squared error is the square difference between the estimates in columns 2a–d and the benchmark estimate of 10.0. MSE is mean squared error, defined as squared error plus the square of the SE. The local linear regression specifications are given in equation (3) in the text. The first two columns present results using bandwidths selected by cross-validation; the last column presents results using the bandwidth calculated by the Imbens-Kalyanaraman method. The bandwidth on each side of the cutoff is given by *h*, in years of age. An *h* of 2.5 means that the sample encompasses those age 52.5–57.5.

*$p < .10$, **$p < .05$ (two-tailed test).

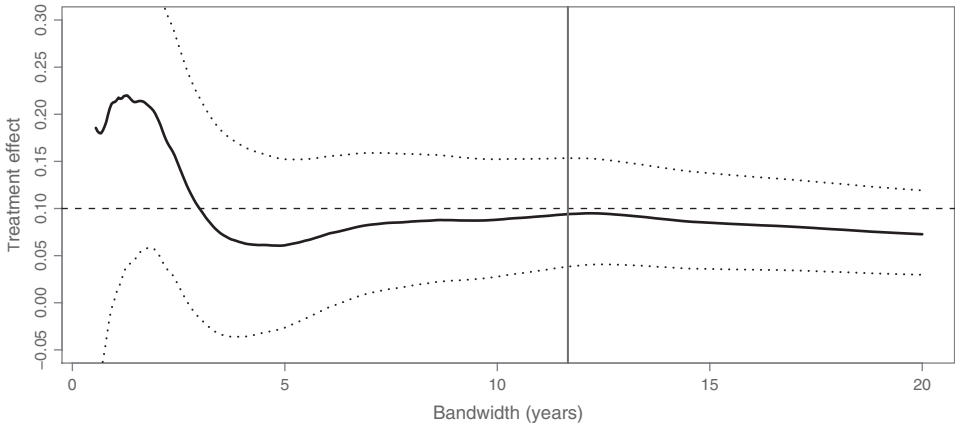

Fig. 3 Local linear regression estimates and confidence intervals, by bandwidth. The graph shows local linear regression estimates of the treatment effect, using various age bandwidths between 0.5 and 20 years. Dotted lines denote 95% CIs. The horizontal dashed line denotes the experimental benchmark estimate from Table 1, column d ($\hat{\beta}_1 = 10.0$). The vertical line denotes the Imbens-Kalyanaraman estimate of optimal bandwidth.

was obviously a poor criterion, although doubtless a tempting one in actual practice. The more systematic procedures associated with local linear regression seem to recommend that approach, especially when the bandwidth size is determined using the Imbens-Kalyanaraman algorithm.

In the simulated example considered above, local regression using optimal bandwidth performs well, but this method is not foolproof. Table 4 traces the success with which this RD method recovers the experimental benchmark when the simulated age cutoff is placed at age 40, 50, 60, and 70. In two of the simulations, local linear regression with optimal bandwidth successfully recovers the benchmark. But in two others, RD misses the benchmark rather badly. In one of the four simulations, the confidence interval that surrounds the RD estimate does not include the benchmark; the 95% interval surrounding the RD estimate based on a discontinuity at age 60 extends from −1.9 to 8.5, but the experimental benchmark is located at 9.9. Looking back on Fig. 1, we see that the relationship between age and turnout is curvilinear near age 60, causing the local linear regression to break down. One could estimate more complex local regression models instead of the linear model used here, but again, this introduces an additional layer of specification uncertainty. Adding covariates (dummy variables indicating whether each person voted in the previous five elections) does nothing to improve the estimate: the estimated effect is 2.5 with a robust SE of 2.6. Bear in mind that the analyst looking at results akin to those presented in Table 4 will not have the luxury of knowing the experimental benchmark and could easily interpret these results to mean that the RD estimate and SE are robust and trustworthy.

Although RD analysis represents an attractive alternative to experiments, one should not lose sight of the fundamental difference between them, a difference that is obscured by reporting conventions. The typical RD analysis reports SEs that account solely for sampling variability. As Gerber, Green, and Kaplan (2004) argue, however, observational approaches such as RD analysis generate estimates whose mean squared error is a function of both sampling variability and specification uncertainty. Experimental and RD estimates may be reported as having the same SEs, but the experimental estimates in such cases have smaller effective SEs because they involve less specification uncertainty.

Table 4 Comparison between experimental benchmarks and local regression estimates using optimal bandwidth, for various simulated age thresholds

	Location of simulated age discontinuity							
	40		*50*		*60*		*70*	
	Benchmark	*RD*	*Benchmark*	*RD*	*Benchmark*	*RD*	*Benchmark*	*RD*
Neighbors treatment (robust SE)	10.68** (1.42)	9.82** (3.85)	10.20** (1.04)	8.08** (2.56)	9.88** (1.22)	3.30 (2.65)	9.63** (1.37)	15.96** (3.85)
Bandwidth (h)		11.97		15.24		15.78		13.10
N (unweighted)	30,038	8419	30,038	10,727	30,038	8031	30,038	4660
N with weights >0		4163		5822		4340		2398
Squared error		0.74		4.49		43.30		40.07
MSE		15.56		11.04		50.32		54.89

Note. Benchmark estimates are based on a fourth-order polynomial regression, as in equation (2) in the text. The local linear regression specifications are given in equation (3) in the text. Squared error is the squared difference between the RD estimates and the corresponding benchmark estimates. MSE is mean squared error, defined as squared error plus the square of the SE of each of the RD estimates.

The challenge for researchers is to reduce this specification uncertainty by building a knowledge base from empirical applications of RD that in turn facilitates the automation of specification and bandwidth selection. Voter mobilization work is ideal for this type of exercise because it involves large samples, frequent use of random assignment, and discontinuities that arise in the ordinary course of political consulting work. But in truth, many data sets are suitable for this type of benchmark exercise, and it is important to canvass a wide array of different applications to assess the conditions under which RD generates reliable estimates.

It should be stressed that benchmarking is also possible using data sets where no intervention has occurred. In the absence of an intervention, the true treatment effect is by construction zero everywhere, obviating the need to estimate a benchmark near the vicinity of the discontinuity. This type of data set may be used to simulate a series of RD analyses with an array of different forcing variables and thresholds. Research questions include (1) the correspondence between the empirical sampling distribution of the RD estimates that emerge from a given estimation procedure (presumably centered around zero) and the sampling distribution implied by the estimated SEs associated with the treatment effect, (2) the relative performance of models that include more/less flexible control functions and wider/narrower bandwidths, and (3) the degree to which diagnostic tools successfully guide the specification of RD models and bandwidths.

By comparing RD estimates to experimental benchmarks, social scientists aim to develop and refine a set of standard operating procedures that can be used reliably when researchers proceed without experimental benchmarks. There is no telling whether the experience acquired through this approach will succeed in generating a useful set of standard procedures, but this line of exploration is nevertheless worth pursuing even at the risk of failure. Researchers currently lack a clear sense of how to gauge the reliability of RD estimates, and the proper quantification of uncertainty is an essential aspect of scientific inquiry.

Funding

Institution for Social and Policy Studies, Yale University.

References

Black, D., J. Galdo, and J. C. Smith. 2005. *Evaluating the regression discontinuity design using experimental data.* Unpublished working paper.

Buddelmeyer, H., and E. Skoufias. 2003. *An evaluation of the performance of regression discontinuity design on PROGRESA.* Discussion Paper Series No. 827. Bonn, Germany: IZA.

Butler, Daniel M., and Matthew J. Butler. 2006. Splitting the difference? Causal inference and theories of split-party delegations. *Political Analysis* 14:439–55.

Cook, Thomas D., and Vivian C. Wong. 2009. Empirical tests of the validity of the regression-discontinuity design. *Annales d'Economie et de Statistique* Forthcoming.

Gerber, Alan, Daniel Kessler, and Marc Meredith. 2009. *The persuasive effects of direct mail: A regression discontinuity approach.* Presented at the 2009 Midwest Political Science Association Meeting.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. Social pressure and voter turnout: evidence from a large-scale field experiment. *American Political Science Review* 102:33–48.

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2004. The illusion of learning from observational research. In *Problems and methods in the study of politics*, eds. Ian Shapiro, Rogers Smith, and Tarek Massoud, 251–73. New York: Cambridge University Press.

Green, Donald P., Terence Y. Leong, Alan S. Gerber, and Christopher W. Larimer. 2008. *Testing the accuracy of regression discontinuity analysis using an experimental benchmark.* Unpublished manuscript, Institution for Social and Policy Studies, Yale University.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. Identification and estimation of treatment effects with a regression discontinuity design. *Econometrica* 69:201–9.

Hainmueller, Jens, and Holger Lutz Kern. 2008. Incumbency as a source of spillover effects in mixed electoral systems: Evidence from a regression-discontinuity design. *Electoral Studies* 27:213–27.

Imbens, Guido, and Karthik Kalyanaraman. 2009. *Optimal bandwidth choice for the regression discontinuity estimator.* Unpublished manuscript, Department of Economics, Harvard University.

Imbens, Guido W., and Thomas Lemieux. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142:615–35.

LaLonde, Robert J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76:604–20.

Lee, David S. 2008. Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics* 142:675–97.

Lee, David S., and Thomas Lemieux. 2009. *Regression discontinuity designs in economics.* National Bureau of Economic Research Working Paper 14723, Cambridge, MA.

Loader, Clive. 1999. *Local regression and likelihood.* New York: Springer.

Ludwig, J., and D. L. Miller. 2007. Does head start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics* 122:159–208.

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

Nickerson, David W. 2007. *An evaluation of regression discontinuity techniques using experiments as a benchmark.* Poster presented at the Annual Meeting of the Society of Political Methodology, July 18–21, State College, PA.

Pettersson-Lidbom, Per. 2004. *Does the size of the legislature affect the size of government? Evidence from two natural experiments.* Unpublished manuscript, Department of Economics, Stockholm University.

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 66:688–701.

Thistlethwaite, Donald L., and Donald T. Campbell. 1960. Regression-discontinuity analysis: an alternative to the ex-post facto experiment. *Journal of Educational Psychology* 51:309–17.