Political Analysis Advance Access published January 8, 2013

Political Analysis (2013) pp. 1–24
doi:10.1093/pan/mps041

How Much Is Minnesota Like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data

Luke Keele
Department of Political Science, Penn State University, 211 Pond Lab, University Park, PA 16802
e-mail: ljk20@psu.edu (corresponding author)

William Minozzi
Department of Political Science, Ohio State University, 2189 Derby Hall, Columbus, OH 43210
e-mail: minozzi.1@osu.edu

Edited by Jonathan Katz

Political scientists are often interested in estimating causal effects. Identification of causal estimates with observational data invariably requires strong untestable assumptions. Here, we outline a number of the assumptions used in the extant empirical literature. We argue that these assumptions require careful evaluation within the context of specific applications. To that end, we present an empirical case study on the effect of Election Day Registration (EDR) on turnout. We show how different identification assumptions lead to different answers, and that many of the standard assumptions used are implausible. Specifically, we show that EDR likely had negligible effects in the states of Minnesota and Wisconsin. We conclude with an argument for stronger research designs.

In recent years, there has been a renewed interest in methods for estimating causal effects using observational data. This interest has led to a greater focus on the assumptions needed for various statistical estimators to produce estimates that can be interpreted as causal. Often, however, assumptions are mistaken for estimators. For example, some assume matching can yield estimates of causal effects, when matching estimators rely on the same specification assumption as regression models.[1]

In this essay, we focus on the role of assumptions in the estimation of causal effects. We start with an outline of the key assumptions behind a number of popular approaches to the statistical estimation of causal effects using observational data. We begin with a discussion of the approaches that depends on the specification of a statistical model. Here, we outline the key assumption needed to make causal inferences based on estimates from regression models, matching estimators, and the differences-in-differences (DID) estimator. We also describe some basic methods for probing the specification assumptions needed for these approaches. Next, we highlight the partial identification approach, where one uses weak, but credible, assumptions but can only bound the causal effect estimate. We then focus on two methods for natural experiments: instrumental variables (IVs) and regression discontinuity (RD) designs.

Next, we present an empirical case study of the effect of Election Day Registration (EDR) in Minnesota and Wisconsin. We argue that the quality of any assumption is hard to assess outside

Authors' note: We thank Michael Hanmer and Devin Caughey for generously sharing code and data. For comments and suggestions, we thank Mike Alvarez, Curt Signorino, Shigeo Hirano, Robert Erikson, Mike Ting, Walter Mebane, Michael Hanmer, Betsy Sinclair, Jonathan Nagler, Don Green, Rocío Titiunik, and seminar participants at Columbia University, the University of Michigan, and the University of Rochester. A previous version of this article was presented at the 2010 Annual Meeting of the Society of Political Methodology, Iowa City, IA, and APSA 2010. Replication files and information can be found in Keele (2012).

[1] See Barabas (2004) for one example of this confusion.

the context of a specific empirical application. This application allows us to closely examine the plausibility of the assumptions needed for each approach. We apply the various approaches and demonstrate how different assumptions lead to different conclusions. We also use a set of techniques to demonstrate that some estimates may be an artifact of the strong assumptions needed for identification. For the methods with the weakest assumptions, we find that there is little evidence to indicate that EDR increased turnout in these two states.

1 Assumptions and Identification

We start with two preliminary tasks. First, we outline notation with the potential outcomes framework (see, e.g., Rubin 1974). The potential outcomes framework, often referred to as the Rubin Causal Model (Holland 1986), has come to be an important tool for understanding the assumptions needed for the estimation of causal effects in both experimental and observational settings. In the potential outcomes model, each individual has two potential outcomes but only one actual outcome. Potential outcomes represent individual behavior in the presence and the absence of a treatment, and the observed outcome depends on the realized treatment status. We denote a binary treatment status indicator with $D \in \{0,1\}$. While D can take on many values, we focus on the binary case for clarity.[2] The potential outcomes are Y_D, and the actual outcome is a function of treatment assignment and potential outcomes, such that $Y = DY_1 + (1 - D)Y_0$.

The potential outcomes framework formalizes the idea that the individual-level causal effect of a law is unobservable, which is sometimes called the *fundamental problem of causal inference* (Holland 1986). We instead focus on population-level estimands, such as the average causal effect, which is $\tau = E[Y_1] - E[Y_0]$. Limits to credible inferences about such causal estimands come in at least two varieties (Manski 2007). First, there are statistical limits. For example, sampling variability limits the conclusions that one can draw based on a small sample of observations. While the statistical problem is obviously important, we concentrate on a second limit to causal inference: the identification problem.

Formally, we observe $E[Y_1|D = 1]$ and $E[Y_0|D = 0]$ instead of $E[Y_1]$ and $E[Y_0]$. Using the law of iterated expectations, the potential outcomes can be decomposed $E[Y_1] = P(D = 1)E[Y_1|D = 1] + P(D = 0)E[Y_1|D = 0]$ and $E[Y_0] = P(D = 1)E[Y_0|D = 1] + P(D = 0)E[Y_0|D = 0]$, where $P\ (D)$ is the population fraction who have a particular value for D. We observe neither $E[Y_0|D = 1]$ nor $E[Y_1|D = 0]$.[3] Therefore, an *identification problem* exists because there are terms in the causal estimand that are not observable. Even if we had unlimited random samples that perfectly represent the population of interest, we still could not estimate the average causal effect without observing both potential outcomes. Resolution of the problem requires an *identification strategy*: a set of assumptions that warrant inferences based on observable quantities. Any research design based on observational data, at least implicitly, adopts an identification strategy. Identification assumptions thus bridge theoretical and observable quantities. When identification assumptions hold, our estimate of the causal parameter is said to be identified, which implies that the confidence interval for the estimated parameter shrinks to a single point as the sample size increases to infinity.

How does identification relate to the fundamental problem of causal inference? An identification strategy is the articulation of an assumption that identifies a proper counterfactual for the treated outcome. Choosing an identification strategy is a choice about what observed quantity is a good counterfactual for the treated units. We now review the various assumptions one might use for identification. All have been used in various parts of the political science literature. In each case, we focus on the assumption most critical for identification.

[2] We also omit a subscript i that would indicate that these are individual-level variables.

[3] Here, we implicitly invoke the stable unit treatment value assumption (SUTVA) through our notation, which permits the assumption that we are actually observing the potential outcomes associated with each treatment condition. While a more detailed discussion of the applicability of SUTVA is outside the scope of the current paper, we do briefly note that SUTVA requires noninterference between treated and untreated units.

1.1 *Standard Operating Procedure: Cross-Sectional Specification Assumptions*

The most common approach to identification is to assume that we observe all relevant covariates, such that the treated and the control groups are comparable to the point that the only reason they differ is that one group received the treatment. The control group under this strategy is a good counterfactual since we observe and adjust for all the covariates that make these units different from the treated group. Critically, this assumption is nonrefutable, insofar as it cannot be verified with observed data (Manski 2007). Under this approach, analysts collect all known confounders and use a statistical estimator to make treated and control groups comparable before the treatment effect is estimated.

The term "selection on observables" is one name for this assumption to emphasize that analysts must observe *all* the covariates that predict both the outcome and the treatment status (Barnow, Cain, and Goldberger 1980).[4] This is not the only requisite assumption for this approach, but it is the critical and untestable assumption needed for identification. Under this strategy, analysts often use a regression model or a matching estimator. In both cases, the researcher must assume that the statistical model is "correctly" specified. While matching estimators rely on a weaker functional form than regression models, matching cannot correct for an "incorrect" specification.

Many argue that specification assumptions either generally do not hold or there is no way to know if they hold in most applications with observational data (Green and Gerber 2002). If one is unwilling to proceed based on a standard specification assumption, what alternatives exist? We now explore a number of alternatives. In each case, we cannot avoid assumptions; it is just that different assumptions are made.

1.2 *Temporal Specification: DID*

The next approach, DID, exploits longitudinal variation to alter the specification assumption in a fundamental way. At first, DID might seem to dominate cross-sectional analysis of treatment effects, but it too relies on an untestable assumption. We need some additional notation; let $t \in \{0,1\}$ indicate time, where $t = 0$ before the treatment is administered, and $t = 1$ after. We now write the potential outcomes as $Y_D(t)$, and the causal effect is now $Y_1(t) - Y_0(t)$. In the simplest case, treatment is only administered after period $t = 0$, so we denote treatment as simply D without respect to time. We write the outcome as $Y(1) = Y_0(1)(1 - D) + Y_1(1)D$. The DID estimand is

$$\begin{aligned}\tau_{\mathrm{DID}} = E[Y_1(1) - Y_0(1)|D = 1] &= \{E[Y(1)|D = 1] - E[Y(1)|D = 0]\} \\ &\quad - \{E[Y(0)|D = 1] - E[Y(0)|D = 0]\}.\end{aligned}$$

The DID estimand is the treated–control difference after treatment adjusted by the treated–control difference from before the treatment. Rather than modeling treatment assignment, DID eliminates treated-control differences across units that are non-time-varying.[5]

What must we assume for τ_{DID} to be identified? Identification, here, requires the expected potential outcomes for treated and control units to follow parallel paths in the absence of treatment. Under this approach, we must assume that the observed and the unobserved differences between the treated and the control groups are constant with respect to time. The control group is a good counterfactual since these units do not change over time, whereas the treatment group changes over time only because of the treatment. This "parallel paths" assumption is also nonrefutable. If there are covariates that predict deviations from a parallel path, these can also be incorporated into a

[4] Other names for this assumption include "conditional ignorability" and "ignorable treatment assignment."

[5] The most straightforward estimation method for the DID treatment effect is to use conditional sample moments, but regression also suffices. Let μ_{dt} be the conditional sample moment for group d in time t. The DID estimate is $\hat{\tau}_{DID} = (\mu_{11} - \mu_{01}) - (\mu_{10} - \mu_{00})$. This quantity can also be estimated with least squares. Let i represent a particular citizen. Then one estimates the linear model $Y_{it} = \beta_0 + \beta_1 t + \beta_2 D_i + \beta_3 t \times D_i + \epsilon_{it}$. Abadie (2005) shows that *plim* $\beta_3 = (\mu_{11} - \mu_{01}) - (\mu_{10} - \mu_{00})$.

statistical model. Again one needs the "correct specification," but now one needs all the relevant covariates that predict the temporal paths of the treated and the control groups. Thus, DID also relies on a specification assumption.

1.3 *Addressing Bias from Unobserved Confounders*

The next approach acknowledges that causal inferences based on specification assumptions are often not credible due to hidden confounders. The response is to develop techniques that address this limitation. We outline three techniques that may clarify whether an association estimated under a specification assumption might be causal or instead reflects a hidden bias due to an unobserved confounder.

Mill (1867) emphasized the need to ensure that treated and control units were identical in all respects save treatment. Fisher (1935, 18) later dismissed this goal as "totally impossible" and advocated random assignment to generate comparable treated and control groups. In an observational study, however, it is often useful to restrict the analysis to a more homogeneous subset of the available data, which can reduce sensitivity to biases from unobserved confounders. This strategy does imply the use of a smaller sample that can lead to imprecise estimates, but uncertainty due to unobserved confounders is far greater in magnitude than sampling uncertainty (Rosenbaum 2005a). In observational data, increasing the sample size limits sampling variability but does nothing to reduce sensitivity to unobserved bias. As Rosenbaum (2010, 102) notes, increasing the sample size with heterogeneous units may increase precision around a biased point estimate, potentially excluding the true effect from the confidence interval. In many cases, a specification assumption may be invoked, but the analyst can opt to use a more homogeneous subset of the data to eliminate heterogeneity. As we demonstrate later, some natural experiments reduce heterogeneity in a formal way.

Cook and Shadish (1994, 95) write, "Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations." Rosenbaum (2005b) develops this idea more formally into the concept of *pattern specificity*, where one uses a pattern of confirmatory tests rather than relying on a single test. For example, comparing the treatment group to different control groups can illuminate the role of unobserved covariates, if the unobservables are thought to differ across the control groups. If a common pattern of effects emerges, the effects are more credibly due to the treatment. Second, causal theories do more than predict the presence of an effect; they also predict the absence of an effect in the absence of treatment (Rosenbaum 2002b). For example, treatment effects detected before a treatment occurred imply clear differences before treatment, and cast doubt that observed differences are causal. Combining a specification assumption with a series of confirmatory tests according to a specific causal pattern may be enough to convince an audience that an estimated association deserves a causal interpretation.

Finally, one can evaluate the robustness of our inferences by conducting a sensitivity analysis. In a sensitivity analysis, we quantify the exact degree to which the identification assumption must be violated in order for our inference to be reversed. Although such analyses are not currently a routine part of statistical practice in political science, they are powerful tools for understanding the magnitude of possible hidden confounders. There are standard methods of sensitivity analysis for the specification assumption (Rosenbaum 2002a; Imbens 2003). We demonstrate many of these techniques in the empirical example that follows below.

1.4 *Partial Identification and Bounds*

Most identification strategies produce *point* identification—a single parameter describes the causal effect. A more radical approach is to abandon point identification. The partial identification approach instead focuses on producing a range of estimates that depend only on weak and very credible assumptions. Under the partial identification approach, the analyst acknowledges that there is a fundamental tension between the credibility of assumptions and the strength of conclusions (Manski 1995).

Manski (1990) argues for first using the weakest possible set of assumptions that are based on the evidence from the data alone. Using the weakest set of assumptions produces a set of bounds (called no-assumption bounds) for the estimate of the causal effect. This strategy isolates ranges of values for the unobservable counterfactuals and produces a range for the average causal effect. Instructively, these bounds always bracket zero. In short, without stronger assumptions, one cannot rule out the possibility that there is no effect. The no-assumption bounds are *not* a confidence interval, but an *identification region*. The notion of an identification region is prior to the notion of a confidence interval, which represents statistical uncertainty.

To make the inference informative, one adds assumptions about the nature of treatment response or assignment. The assumptions must be based on substantive insights about the process under study. These additional assumptions narrow the bounds on the treatment effect. By adding the assumptions individually, it allows one to observe exactly which assumption provides an informative inference. Assumptions can also be combined for sharper inferences. Next, we outline two common assumptions often used to narrow no-assumption bounds.

First, we can assume the treatment is not counterproductive, so that it has a monotone response (Manski 1997).[6] This is tantamount to assuming that we know the sign of the average causal effect. This assumption is often referred to as a monotone treatment response (MTR). Independent of any assumption about response, one can make an assumption about assignment to treatment. For example, one can assume monotone treatment selection (MTS), which means that average potential outcomes are higher for individuals under the treatment than for those who do not receive the treatment.[7] Under MTS, we assume that treated units are selected to maximize the outcome. Finally, we could combine both assumptions to narrow the bounds further.

This strategy has three strengths. First, the role of the assumptions in the analysis is completely transparent. Second, the treatment effect estimate can easily be assessed according to the plausibility of the identifying assumption. Finally, we avoid any type of specification assumption. We need not assume that we have correctly specified either a model for the outcome or treatment. In short, we can proceed under very weak assumptions, though this will come at cost, since we can never rule out that there is no effect.

1.5 *Natural Experiments*

The final approach we discuss is that of natural experiments. Some in economics credit natural experiments with having produced a "revolution" in the study of observational data (Angrist and Pischke 2010). We define a natural experiment as a real-world situation, which produces haphazard assignment to a treatment. The hope is that a natural intervention will create as-if randomized treatment assignment and produce the same counterfactual comparison that occurs in a randomized experiment. Of course, randomization in an experiment is a fact, whereas haphazard treatment assignment often requires considerable judgment to justify it as as-if random.

We review two forms of natural experiments: IVs and RD designs. Both of these methods provide an estimate for a subset of the study population. This is an important point: the leverage of both methods is predicated on reducing heterogeneity in the study population by making a more focused comparison. As such, one might view the efforts to reduce heterogeneity outlined in Section 1.3 as an attempt to mimic natural experiments.

1.5.1 IVs

One approach is to find an instrument for treatment status, where an instrument is a random encouragement to accept the treatment. We do not provide a full account of the assumptions needed to identify estimates as causal in the IVs context. Sovey and Green (2011) provide a recent review of these assumptions, and Angrist, Imbens, and Rubin (1996) fully derive these

[6] Monotone response requires $E[Y_1|X,D] \geq E[Y_0|X,D]$ for all x and D.

[7] Monotone selection requires $E[Y_D|1] \geq E[Y_D|0]$ for $D \in \{1,0\}$.

assumptions. Instead, we outline an experimental design where IV is valid, and we use this design as a heuristic for understanding whether the IV approach is valid in the context of an observational study. The IV method is identified in what is known as the randomized encouragement design. Under this experimental design, subjects are randomly encouraged to take a treatment. Some subjects refuse or fail to take the treatment, but the object of inference is the effect of the treatment and not the randomly assigned encouragement. If the assumptions hold, the IV method provides an estimate of the treatment as opposed to the encouragement. The IV estimand is the average effect among those induced to take the treatment by a randomized encouragement known as the complier average causal effect or the local average treatment effect (LATE).

An example is useful. In one classic application of the encouragement design, subjects are randomly encouraged to exercise. Some subjects choose to exercise, whereas others do not. Later, health outcomes are measured for all participants. The IV estimate will identify the effect of exercise as opposed to the effect of encouragement even though not all subjects exercise. In the context of natural experiments, one hopes to find some intervention that randomly encourages units to take the treatment. Often the assumptions behind IV are no more plausible than specification assumptions. Again, careful evaluation of the assumptions within the context of a specific application is critical.

1.5.2 RD Designs

Recently, use of the RD design has been revived in the social sciences. The promise behind RD designs arises from the relatively weak assumption needed to identify treatment effects. Below, we briefly outline the RD design. Interested readers should see Lee and Lemieux (2010) for a detailed introduction.

In an RD design, assignment of a binary treatment, D, is a function of a known covariate, S, usually referred to as the *forcing variable* or the *score*. In the sharp RD design, treatment assignment is a deterministic function of the score, where all units with a score less than a known cutoff in the score, c, are assigned to the control condition ($D=0$), and all units above the cutoff are assigned to the treatment condition ($D=1$). Hahn, Todd, and van der Klaauw (2001) demonstrate that for an RD design to be identified, the potential outcomes must be a *continuous* function of the score. Under this continuity assumption, the potential outcomes can be arbitrarily correlated with the score, so that, for example, people with higher scores might have higher potential gains from treatment.

The continuity assumption is a formal statement of the idea that individuals very close to the cutoff, but on opposite sides of it, are comparable or good counterfactuals for each other. When the continuity assumption holds, treatment assignment close to the cutoff is as-if random. That is, the RD design identifies a LATE for the subpopulation of individuals whose value of the score is at or near c. Estimation of the RD treatment effect proceeds by selecting a subset of units just above and below the discontinuity and calculating the difference across these two groups. As we discuss in Appendix A, there are a number of different methods for selecting units around the threshold as well as for the estimation of the treatment effect.

Lee (2008) provides a behavioral interpretation for the continuity assumption. He writes the score as $S=W+e$, where W represents efforts by agents to sort above and below c and e is a stochastic component. When e is small and agents are able to precisely sort around the threshold, the RD design may not identify the parameter of interest. In this case, treatment is completely determined by self-selection and the potential outcomes will be correlated with observed and unobserved characteristics of the agents. However, when e is larger, agents will have difficulty self-selecting with any precision, and whether an agent is above or below the threshold is essentially random. The behavioral interpretation of the continuity assumption often allows for easier assessment of the identification assumption in the RD design.

Another advantage of the RD is that the identifying assumption has a clearly testable implication. In the RD design, variation in the treatment is approximately random within some local neighborhood around the threshold, and when true, all "baseline characteristics"—all those variables determined prior to the realization of the assignment variable—should have the same distribution just above and below the cutoff. If there is a discontinuity in these baseline covariates, then the underlying identifying assumption in an RD is unwarranted (Lee and Lemieux 2010).

When we have reasons to believe that e is small relative to W, we maintain that the RD design is the strongest of the approaches that we have reviewed. In the RD design, we have a clear testable implication of the identifying assumption. Second, the RD is based on a design. That is, we do not rely on found data, but instead we rely on the fact that a threshold had to be created and implemented. Recently, RD designs have gained further credibility by recovering experimental benchmarks (Cook, Shadish, and Wong 2008)

There is a drawback to the RD identification strategy. The RD estimate is necessarily local: it only applies to some limited range of units above and below the threshold, but not to all units. One goal of our analyses might be to make policy recommendations, and a local treatment effect by definition may not extrapolate to other units. With observational data, however, we must often trade external validity for internal validity. In this sense, the RD embodies the call for more local estimates that we outlined in Section 1.3.

2 Case Study: EDR

We now turn to an empirical application to highlight the role of assumptions in the estimation of causal effects with observational data.[8] More specifically, we examine the effect of EDR on turnout. EDR significantly reduces the cost of voting by collapsing voting and registration into the same act. EDR is widely credited with increasing turnout (Wolfinger and Rosenstone 1980; Teixeira 1992; Mitchell and Wlezien 1995; Rhine 1995; Highton and Wolfinger 1998; Timpone 1998; Brians and Grofman 1999, 2001; Knack 2001; Hanmer 2007, 2009). Based on this empirical evidence, political scientists are often willing to argue in publications such as the *New York Times* that if all states adopted EDR, turnout would increase nationwide (Just 2011). Thus, we examine a policy intervention that many believe has increased turnout, whose effects are thought to be well understood and that political scientists propose as good policy. The analysis below is not meant to be a comprehensive study of EDR, but it is meant to highlight how inferences differ depending on what assumptions are used for identification.

Below, we focus on EDR in Minnesota and Wisconsin for two reasons. First, these two states are among those with the highest turnout, and EDR has long been credited with this achievement (Wolfinger and Rosenstone 1980). Second, while some work has cast doubt on whether EDR increased turnout in states like New Hampshire, Wyoming, and Montana, it is still widely understood to have increased turnout in Minnesota and Wisconsin (Hanmer 2009). Therefore, we use Minnesota and Wisconsin as our treated states. This implies that turnout in the other forty-eight states must serve as our counterfactual. The difficulty is that other states may have different levels of turnout for many reasons other than the fact that they do not have EDR. In each approach below, we will have to rely on assumptions to create a valid counterfactual comparison. In the later analyses, we narrow the scope of the analysis and compare voters within a single state. More importantly, we restrict the counterfactual comparison even further by comparing cities of nearly equal population.

We start with a bounds analysis to understand what we can learn with minimal assumptions. Next, we use both the cross-sectional and temporal specification identification strategies. We then examine these estimates for bias due to unobserved confounders. We conclude with estimates from an RD design.

2.1 *Manski Bounds*

We start with a bounds analysis to understand what we can learn from the data without strong assumptions.[9] In the bounds analysis, we compare turnout in Minnesota and Wisconsin, the two treated states, to turnout in the rest of the country. The counterfactual in this analysis is all

[8] Full replication files are available in Keele (2012).

[9] For examples of more complete analyses based on bounds, see Hanmer (2007, 2009).

Table 1 Manski bounds analysis for the effect of EDR in Wisconsin and Minnesota, 1976 and 1980

	1976	*1980*
No assumption bounds	−61 to 39	−63 to 37
MTR	0 to 39	0 to 37
MTS	−61 to 18	−63 to 12.7
MTR + MTS	0 to 18	0 to 12.7

Note. Each set of bounds represents a lower and upper bound on the EDR treatment effect in percentage terms.

available voters outside the two treated states. The advantage with the bounds approach is that we rely on a set of very weak assumptions about the comparability of the control voters to the treated voters. The disadvantage is that bounds leave us with a great deal of uncertainty about whether EDR was actually effective. Minnesota first used EDR in 1974, whereas Wisconsin first used it in 1976. We calculate the bounds using data from both 1976 and 1980. We use data from 1980 to allow for a possible delay in the onset of the EDR effect.

We start with the no-assumption bounds in the first row of Table 1. Recall that these are bounds on the identification region and do not reflect any statistical uncertainty.[10] These bounds reveal what we can learn from the data alone without *any* assumptions about identification. In 1980, the no-assumption bounds range from −63 to 37. That is, without any assumptions, we can say the effect of EDR ranges from depressing turnout by 63% to increasing it by 37%. It is instructive to see how little can be inferred without assumptions.

To make the bounds more sharp, we must add assumptions. We next assume that monotonicity holds. This is tantamount to assuming that we know the sign of the treatment effect. In the context here, the MTR assumption implies that EDR does not depress turnout. MTR is a fairly weak assumption in this context, since it is difficult to imagine how EDR would cause a citizen to not vote. The MTR bounds are in the second row of Table 1. Under this assumption, in 1980, EDR raised turnout by as much as 37% or as little as zero.

We next calculate the bounds assuming MTS, or that treated units are selected to maximize the outcome. That is, we assume that Minnesota and Wisconsin enacted EDR under the assumption that it would increase turnout. In the context of EDR, MTS is questionable. In Minnesota, EDR served as a compromise when policymakers wanted to implement a statewide voter registration system, and some areas of the state resisted since they did not have a registration system (Smolka 1977). Thus, EDR was the result of a legislative compromise and not an outburst of civic engagement. The third row of Table 1 contains the bounds calculated under MTS. Under this assumption, in 1980, the EDR effect ranges from −63% to 12.7%.

If we now combine MTS and MTR, the range of the EDR effect in 1980 narrows to 0% and 12.7%, respectively. In 1976, these bounds range from 0% to 18%, which indicates that the EDR effect perhaps faded instead of growing over time. Thus, under a fairly weak set of assumptions, we can infer that EDR increased turnout by as much as nearly thirteen points in 1980, but that effect may also have been zero or anywhere in between. The cost of weak assumptions is a range of estimates and consequently greater uncertainty about the true treatment effect. We now estimate the treatment effect under specification assumptions, which will provide what appears to be greater certainty but at the cost of much stronger untestable assumptions.

2.2 *Specification Approaches*

Many analysts use a cross-sectional specification approach to causal inferences about turnout. That is, they use a regression model to adjust for differences in state-level distributions of education, income, and other similar covariates. In doing so, they implicitly assume that they observe all the

[10] The bootstrap may be used to provide estimates of statistical uncertainty. The sample sizes are large, so sampling uncertainty is small here.

Table 2 Logistic regression estimate of the effect of EDR in 1976 and 1980

	1976	*1980*
Treatment effect estimate (95% CI)	12.5 (10.2–14.8)	10.1 (8.7–11.7)

Note. Estimate from logistic regression. Confidence intervals are calculated using the bootstrap. Cell entries are in percentages. Estimate is Minnesota and Wisconsin compared to all other states. Model includes a full specification with measures of education, income, race, sex, marital status, age, and a dummy variable for Southern states.

reasons that voters in states without EDR differ from voters in the two states with EDR. Recasting the specification assumption as a selection on observables assumption helps illuminate the fact that while these models may have a reasonably good specification for turnout, there are few if any covariates that predict why a state adopts EDR. Put another way, the covariates that are correlated with turnout do not provide much leverage over why respondents in one state are treated to EDR and those in another state are not. In short, regression models do not account for why some states select into the EDR treatment and other states do not. Moreover, this approach requires us to make comparisons *across* states. Such comparisons are problematic since cross-state differences in turnout are a function of a wide variety of processes. Things such as electoral competitiveness that can be magnified by the electoral college, specific statewide elections, political culture, or differences in other state election procedures, such as polling hours or absentee balloting, can all contribute to differences in turnout at the state level. Unless we are confident that we have fully measured and controlled for all these varied factors, any attempt to isolate cross-state turnout differences due to EDR will be confounded.

Here, we estimated the effect of EDR in Minnesota and Wisconsin with data from 1976 to 1980. We used a logistic regression model specified with measures for income, education, age, age-squared, a dummy variable for whether the respondent is African American, marital status, sex, a dummy variable for Southern states, and time at address before the election. The results are in Table 2. Under a specification assumption, EDR increased turnout 12.5% points in 1976 and 10.1% points in 1980. Thus, there is some evidence that the EDR effect declined over time. As a point of reference, we estimated the same model with data from 2008, and the estimated EDR effect was seven percentage points. Unlike the bounds approach, now that we have adopted a strong untestable assumption, we get a precise yet dubious estimate that indicates that EDR produced a large increase in turnout in both years.

We next use an alternative specification approach: DID. When applied to EDR, DID has both advantages and disadvantages. First, the DID estimator is clearly superior to the specification approach used with cross-sectional data. The DID estimator will account for baseline differences in turnout across states, which are quite common. However, the DID estimator also has serious drawbacks particularly when applied to turnout. Recall that with the DID estimator, we must assume that the differences between the groups are constant across time absent treatment. Here, we assume that voters in other states are good counterfactuals because the differences in turnout before EDR goes into effect are constant with respect to time. That is, we must assume that no other events beside the treatment alter the temporal path of turnout for either the treated or control groups. In the context of voter turnout, this implies that nothing else in the treated states serves to boost turnout. However, a competitive senatorial or gubernatorial race in the treated state could boost turnout, and the result of such an event would be attributed to EDR. As another example, if a state that has adopted EDR becomes a battleground state in the next election, it will be impossible to distinguish the effects of EDR from increased mobilization efforts that result from increased competition. Again, the standard turnout model specification with measures such as education and income that are nearly constant across elections provides little predictive power for the overtime dynamics of turnout. Thus, while the DID strategy may be plausible in many contexts, this assumption is less plausible in the study of turnout across states, where it is not unusual for other factors to alter the dynamics of turnout across elections.

Table 3 DID estimate of the effect of EDR, 1972–76 and 1972–80

	1972–76	*1972–80*
Treatment effect estimate (95% CI)	5.8 (1.8–9.9)	4.1 (0.76–7.3)

Note. Cell entries are in percentages. Model includes a full specification using education, income, race, sex, and age. Estimate in Minnesota and Wisconsin compared to the rest of the states. Standard errors adjusted for clustering at the state level.

To estimate the treatment effect using DID, we used 1972 as the pretreatment year and 1976 and 1980 as posttreatment years.[11] We use standard errors clustered by state, although this may not be sufficient to yield correct estimates of the standard errors (Donald and Lang 2007). We also use a slightly different empirical specification, including measures of education, sex, a dummy variable for African Americans, income, age, and age-squared, since not all the measures in 1980 or 1976 were available in 1972. Table 3 contains the results from the DID method. Under DID, the EDR effect is now 5.8 points instead of 12.5 in 1976 and 4.1 points instead of 10.1 in 1980.[12] Given that the DID estimates are more than half the size of the cross-sectional estimates, it suggests that the specification approach with cross-sectional data is biased. Much of the EDR effect in that model is due to the myriad differences that cause turnout to differ across states in each election. The DID estimator appears to correct for this, but perhaps not sufficiently.

One advantage of the DID estimator is that one can use the data to perform an informal test of the identifying assumption. To do so, one plots the trend in the treated and the control outcomes before the treatment goes into effect (Angrist and Pischke 2009). If the two trends are largely parallel, then this is the evidence that the assumption holds. Figure 1 contains a plot of the average turnout in Wisconsin and Minnesota when compared to the average turnout in the rest of the United States. Based on Fig. 1, the DID assumption appears to be credible, as the pretreatment trends are similar.

2.3 *Addressing Bias from Unobserved Confounders*

If we relied on a specification assumption in either form, we would conclude that EDR did increase turnout in Minnesota and Wisconsin. We now probe for bias from unobserved confounders. We do this in three ways. First, we attempt to reduce heterogeneity through a more focused comparison. Second, we use a sensitivity analysis for the specification assumption. Third, we examine the data for patterns specific to the causal hypothesis.

Thus far, we have compared voters from the two treated states, Minnesota and Wisconsin, to voters in all the other states. In short, we are using a weighted average of other states as a counterfactual for these two states. An alternate strategy that we highlighted in Section 1.3 is to restrict the study population to the areas of the United States that are more comparable to Minnesota and Wisconsin. Such a restriction should reduce heterogeneity and potentially reduce bias in the estimates. Hanmer (2009) uses this approach by using Iowa and South Dakota as control states in a DID analysis, but neither state has a large metro area comparable to Milwaukee and Minneapolis–St. Paul. This is important since in the current population survey (CPS) sample, comparing Minnesota and Wisconsin to Iowa and South Dakota is to mostly compare urban voters to rural voters. In 1972 and 1976, over 90% of the CPS sample in Minnesota and Wisconsin is drawn from the Minneapolis–St. Paul and Milwaukee metropolitan areas.

[11] Importantly, we have elided the differences between the average treatment effect (ATE) and the ATE on the treated (ATT). In the case study, we estimate the effect of laws that apply to all citizens in a state; we are effectively assuming that the distinction between these quantities is not substantial enough to be substantively relevant. Of course, in many, many other environments, the distinction between ATE and ATT is meaningful. In those cases, the regression estimator (which estimates ATE) and the DID estimator (which estimates ATT) are not directly comparable.

[12] Our estimates are very similar to those in Hanmer (2009).

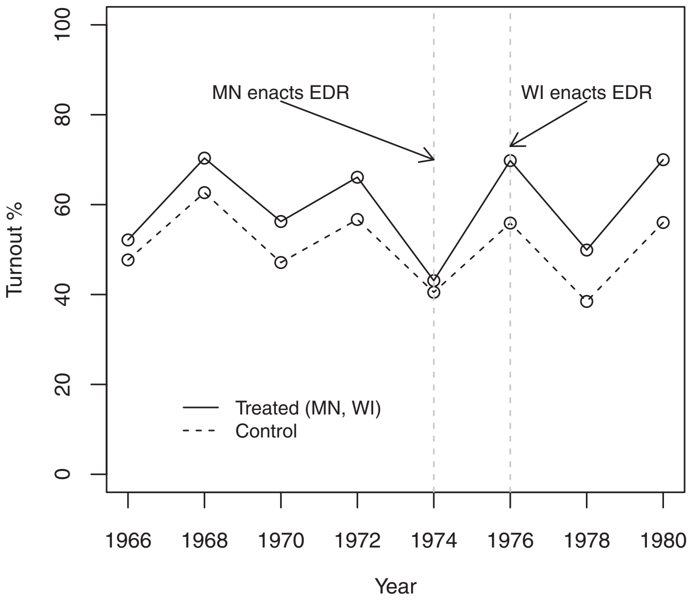

Fig. 1 Trends in presidential election turnout for treated and control states.

Table 4 Difference in turnout rates across Minnesota and Wisconsin, 1974

Unmatched estimate[a] (95% CI)	9.6 (3.9–15.2)
Matched estimate[b] (95% CI)	9.7 (3.9–15.2)
Γ	1.2
N	866

Note. Cell entries are in percentages.
[a]Unmatched estimate is simply a difference in proportions test for all CPS respondents across Minnesota and Wisconsin 1974.
[b]Matched estimate is based on matching respondents across Minnesota and Wisconsin.

To avoid such a comparison, we take advantage of the timing of EDR implementation. Minnesota first used EDR in the 1974 election, whereas Wisconsin did not have EDR in place until 1976. We exploit the timing in two ways. First, we conduct an analysis in 1974, where we compare Minnesota to Wisconsin. This allows for a comparison among the two states that actually implemented EDR first. We performed a matching analysis in 1974, where we matched on the same covariates used in earlier analyses. We rely on matching since it allows us to more easily implement a sensitivity analysis.[13]

In Table 4, we present two different treatment effect estimates for EDR in Minnesota. The first estimate does not adjust for any covariates, whereas the second estimate uses matching to adjust for education, income, age, marital status, and time at address. We excluded a small number of racial minorities from the analysis. We draw attention to two different aspects of the estimates. First, the covariates do little in that the treatment effect estimate barely changes once we match. This demonstrates how much adjustment is provided by simply selecting comparable states.

Second, we perform a sensitivity analysis. Recall that in a sensitivity analysis, we quantify the exact degree to which the identification assumption must be violated in order for our inference to be

[13] We use genetic matching (Sekhon 2011; Sekhon and Diamond forthcoming) to form two essentially equivalent groups based on the observed CPS covariates. Balance was quite good with the smallest p-value from KS tests for balance being above 0.40. Note that for the matching estimator, we actually estimate the ATT instead of the ATE.

changed. We apply a sensitivity analysis for matching estimators developed by Rosenbaum (2002b). In the sensitivity analysis, two subjects with the same observed characteristics may differ in the odds of receiving the treatment by at most a factor of Γ. In a randomized experiment, randomization of the treatment ensures that $\Gamma = 1$; that is, the odds of treatment are the same across the treated and the control groups. In an observational study, Γ may depart from one. For example, if Γ is two for two matched subjects, one treated and one control, that are identical on matched covariates, then one subject is twice as likely as the other to receive the treatment because they differ in terms of an unobserved covariate (Rosenbaum 2005b). While the true value of Γ is unknown, we can try several values of Γ and see how the conclusions of the study change. Specifically, we calculate an upper bound on the p-value for a range of Γ values. If the upper bound on the p-value exceeds the conventional 0.05 threshold, then we conclude that a hidden confounder of that magnitude would explain the observed association. If the study conclusions hold for higher Γ values, we can conclude that the estimate is robust to the presence of a hidden confounder.

The third row of Table 4 contains the value of Γ at which the p-value exceeds 0.05. We see that if an unobserved covariate caused two identically matched voters to differ in their odds of treatment by 1.2, then that would explain the estimated effect. Normally, we would compare this number to the estimated effect sizes of other covariates that predict treatment. However, in our example, the covariates distributions across the treated and the control groups are so similar before matching that estimated odds ratios are too small for useful comparison. To make sense of the magnitude of Γ, we use a technique known as amplification (Rosenbaum and Silber 2009). In the basic template for the sensitivity analysis we outlined above, Γ is the degree of association between u, an unobserved confounder, and D, when u is perfectly correlated with Y, the outcome. In this context, however, we have a model for the outcome but not for the treatment status. Rosenbaum and Silber (2009) show that Γ can be decomposed into a set, known as an amplification set, (Δ,Λ), through the following equation: $\Gamma = (\Delta\Lambda + 1)/(\Delta + \Lambda)$. In the amplification set, the parameter Δ is the association between u and Y, and Λ, which is the degree of association between u and D. The amplification set allows us to recast the sensitivity analysis in terms of Δ, the association between the confounder and the outcome. Therefore, we can understand the degree of confounding needed to explain the observed association in terms of the outcome as opposed to the treatment.

For a given Γ value, there are a number of values of Δ and Λ that are possible. First, we calculated the amplification sets consistent with $\Gamma = 1.2$. Next, we calculated the odds ratios for the covariates in the turnout model for 1980 in Table 1. We then formed an amplification set consistent with the effect sizes in that model. For example, we found that in the logit model for turnout that being married increases the odds of voting by 1.4, which is consistent with an amplification set of (1.4–3.0). This implies that a confounder that triples the odds of treatment within a matched pair and increases the odds of turnout by 1.4 would explain the association we observe. As such, if we failed to control for a confounder that has an effect comparable to being married, then that would explain the observed association. We would argue that being married is a relatively minor and theoretically unimportant factor in the literature on turnout. For example, in the same model, a 4-year increase in education increases the odds of voting by 5.10. Thus, if we failed to control for a relatively minor covariate, like being married, it could reverse our conclusion that EDR increased turnout in MN.

Next, we examine these results in terms of pattern specificity. That is, we ask whether the estimated effects fit a broader pattern that is consistent with the causal hypothesis. First, we perform an informal placebo test. If Wisconsin is a good counterfactual for Minnesota, turnout rates in the two states should be similar before EDR went into effect in Minnesota in 1974. We are unable to perform a more formal placebo test, since the turnout items were not a part of the CPS in 1970, the first midterm election before EDR went into effect in 1974. Instead, we use actual turnout rates for the two states. Table 5 contains turnout rates for Minnesota and Wisconsin and the difference from 1966 to 1980. Importantly, these data reveal that turnout in Minnesota was higher than turnout in Wisconsin in all pretreatment years and that difference was at least six points. This suggests that the effect we estimated in Table 4 is simply an estimate of the fixed difference in turnout between the two states.

Table 5 Turnout rates in Minnesota and Wisconsin, 1966–80

Year	*Minnesota*	*Wisconsin*	*Difference*[a]
1966	57.7	46.6	11.1
1968	73.6	67.1	6.5
1970	60.9	51.6	9.3
1972	69.2	63.0	6.2
1974	46.8	39.4	7.4
1976	72.3	67.3	5.0
1978	54.8	45.0	9.8
1980	71.4	68.6	2.8

Source: *A Statistical History of the American Electorate* (Rusk 2001).
[a]The difference is calculated as the Minnesota turnout rate minus the Wisconsin turnout rate.

We can examine the data in Table 5 to see whether they are consistent with other causal patterns that should be present. As we outlined before, in 1974, Wisconsin did not yet have EDR, whereas Minnesota did. If EDR is effective, the turnout difference in 1974 should vanish or shrink after the treatment goes into effect in Wisconsin in 1976. If, instead, the estimated gap stays roughly constant, it is unlikely that EDR had any effect. This sequencing of the treatment effect pattern relies on a posttreatment comparison, so we must assume that no other interventions alter the trajectory after treatment. It does, however, allow us to look for a pattern of effects that is consistent with our causal hypothesis. Given turnout differences between presidential and midterm elections, we use 1978 as the posttreatment comparison year for 1974. When we compare 1974–78, we expect the difference in turnout to shrink from the estimated more than seven-point gap in 1974. However, the difference in turnout in 1978 is nearly ten points, which is actually larger. As such, the estimated differences do not fit the causal pattern we would expect, where the difference between the two should shrink. Finally, using the data in Table 5, we can also perform an informal DID analysis. Here, we use turnout in 1970 as the pretreatment baseline and use 1974 as the posttreatment period for the DID estimates. This DID estimate is −1.9%, which of course is in the wrong direction.[14]

In sum, the evidence for an EDR effect is not compelling once we probe for evidence of hidden confounders. The matching analysis reveals how weak the specification is once we restrict the analysis to comparable states. The sensitivity analysis suggests a hidden confounder could easily explain the estimated difference. Finally, we see that the general turnout pattern is not consistent with the causal hypothesis. Perhaps the problem is that our estimates are not local enough. Next, we analyze natural experiments in both Wisconsin and Minnesota.

2.4 *Natural Experimental Approach: RD*

We use natural experiments to more closely approximate the counterfactuals produced by a randomized experiment. First, we ask whether we can find an appropriate instrument to allow for IV estimates of the treatment effect. To evaluate IV in this context, we must ask how well EDR fits the paradigm of the randomized encouragement design. We argue that the fit is poor. In the case of EDR, we must find an instrument that randomly encourages states to adopt EDR but has no subsequent direct effect on turnout. We know of no instrument that fits these criteria. While the IV approach may be successful in some venues, we argue that it tends to be implausible for studying interventions like EDR.

Is it possible to conduct an RD analysis in the context of EDR? Yes, we can for both states. Before the adoption of EDR in Wisconsin and Minnesota, registration systems were based on municipal population. In Wisconsin, municipalities with populations of less than 5000 were not

[14]We plotted the turnout rates, and the trend is quite similar in the pretreatment time periods.

required to use a voter registration system, whereas those municipalities with populations greater than 5000 had a standard system of registration (Smolka 1977). In Minnesota, the same system was in place but the population threshold was 10,000. In 1976, once EDR was adopted in Wisconsin, municipalities with populations of less than 5000 were still not required to maintain a registration system, but those with a population greater than 5000 switched to an EDR registration system. In Minnesota, all municipalities adopted the EDR system regardless of population.[15] See Appendix A for details on how we collected the data for this analysis.

The RD design has two distinct advantages in the context of EDR. First, we can implement the design within a single state, which holds all state-level confounders constant without specification assumptions. The key difficulty with the other approaches is that we are unable to provide a compelling specification for why turnout differs a cross states. The best we have been able to do thus far is to rely on state fixed effects with the DID estimator. Given that the turnout may differ across states due to a variety of institutions, history, and competitiveness, the best approach may be to hold all state-level confounders factors constant using a within-state design. We are not the first to adopt a within-state design. Other studies have used the DID approach within a single state (Ansolabehere and Konisky 2006; Burden and Neiheisel 2011b). These studies tend to find effects that are much smaller than most in the literature.

Second, the RD design is credible in this setting, since we expect that it will be difficult for municipalities to manipulate their population in a precise manner to avoid having to use a registration system. Population for municipalities is determined by the US Census in 1970 and not the municipality, so it should be nearly random whether a municipality has a population either just above or below 5000 or 10,000 in each state.

The RD design comes in both a sharp and a fuzzy form. We have some reason to believe the fuzzy version might be appropriate in Wisconsin. In the fuzzy version of the RD, other factors besides the threshold affect the probability of receiving the treatment. For a fuzzy RD design, the assignment to treatment is a random variable given the score, but the probability of receiving treatment conditional on the score, $P(D = 1|S)$, still jumps discontinuously at c. This implies that it is possible for some units with scores below c to receive the treatment. The fuzzy RD design results in an equivalence between the RD and the IV design (Hahn, Todd, and van der Klaauw 2001). As such, the fuzzy RD requires additional assumptions.[16]

Why do we suspect a fuzzy RD design might be appropriate in Wisconsin? Some municipalities below the population threshold began using a registration system voluntarily before EDR went into effect, based on approval by local voters (Smolka 1977). We do not have any evidence on how widespread this was or which municipalities may have adopted a registration system. Thus, compliance below threshold may not have been perfect. Imbens and Kalyanaraman (2012) develop an estimation method for both the sharp and the fuzzy design, and we estimate the RD effect both ways.

There is one key limitation associated with the RD design in this context: in both states, RD identifies an effect but does not identify the effect of interest. Ideally, we would observe an RD design where municipalities below 5000 or 10,000 use a standard voter registration system and those above the threshold have EDR or vice versa. This design would isolate the precise EDR effect. What we actually observe is that municipalities below the threshold have no registration when compared to municipalities with either a standard registration system or EDR. It is not unusual for a natural experiment to identify an effect, but the effect that is identified may be subtly different from the effect of interest. Sekhon and Titiunik (2012) provide an excellent case study of this phenomenon. Does this mean all is lost?

In Wisconsin, we could conduct two separate RD analyses, the first before the adoption of EDR and the second after. If EDR has an effect, the gap in the treatment effect between these two RD

[15] After the adoption of the statewide registration system and EDR, counties with no municipalities larger than 10,000 were allowed to be exempt from registration. Only a single county, Pope County, chose to be exempt (Smolka 1977). This does not affect the analysis we outline below.

[16] Specifically, one must assume that the exclusion restriction and monotonicity hold.

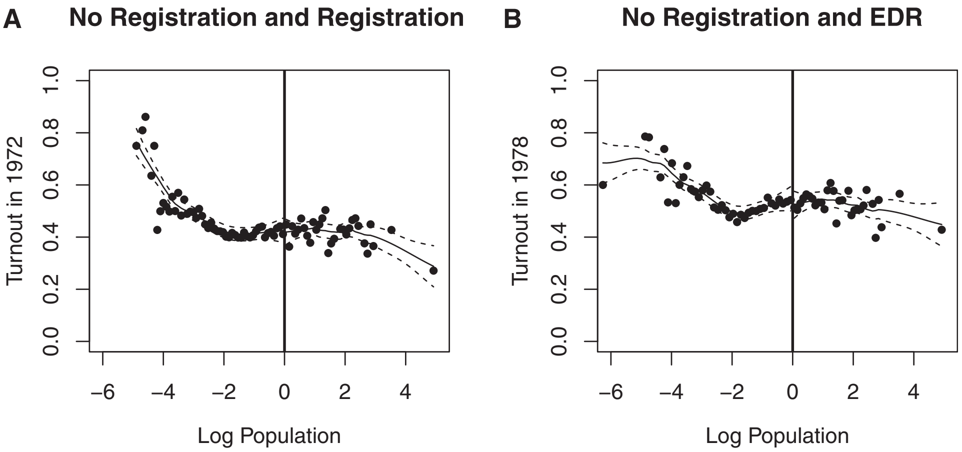

Fig. 2 RD design: the effect of voter registration system on turnout in Wisconsin. (**A**) No registration and registration and (**B**) No registration and EDR.

estimates should shrink. This is essentially a mixture of RD and DID.[17] Comparing the two RD estimates, however, requires adopting the DID assumptions which, as we have argued, may not be realistic. Moreover, it seems imprudent to adopt one identification strategy due to its weaker assumptions and then add another, more implausible, assumption. This strategy is not possible in Minnesota since the entire state adopted EDR.

We argue that a better approach is to use the RD estimate from 1974 in Wisconsin and from 1972 in Minnesota as an upper bound on the effect of interest.[18] That is, the gap between no registration and registration should form an upper bound on the EDR effect. For example, in Wisconsin, if turnout is five points lower for municipalities with a population of 5000 or more before EDR goes into effect, we cannot expect EDR to be larger than five points. Moreover, if there is no difference in turnout before EDR goes into effect, we cannot credibly argue that EDR increases turnout. Thus, despite the fact that we cannot identify the effect of interest, we can at least put an upper bound on the EDR effect. For Wisconsin, we also estimate the RD treatment effect for 1978 but only as a means of informal comparison.

One important strength of the RD design is that the identification assumption has a testable implication. If we apply the RD analysis to other covariates that should be correlated with turnout, we should not find any effect. We tested for discontinuities with eight important census covariates aggregated to the municipal level: percentage African American, percentage over the age of 65, percentage with a high school degree, percentage below poverty, per capita income, median household income, median house value, and median rent for the subset of municipalities in Wisconsin and Minnesota that had census data. We plotted each of these covariates against log population along with a nonparametric regression fit to both sides of the discontinuity. Importantly, there is no evidence that any of these covariates are correlated with the score near the threshold (see Appendix A). We also performed one additional diagnostic. We implemented the density test suggested by McCrary (2008) to check for evidence of manipulation in the forcing variable. This test is analogous to checking whether the ratio of treated to control units in an

[17] Burden and Neiheisel (2011a) exploit the same variation in Wisconsin voting laws, but apply the DID estimator. They find a positive effect for EDR but only when they include Milwaukee. As such, their estimate is far less local than that in an RD design. We found if one estimates the RD effect estimate with all the data, Milwaukee exerts a strong influence. This is to be expected, but Milwaukee has no reasonable counterfactual within the state of Wisconsin. In general, we think a DID estimator here fails to exploit the much weaker assumptions in the RD design.

[18] One might argue that the effect might differ in a presidential election year. One could also argue that the effect might differ when an election is more competitive and voters are more willing to pay the cost of registration. A more fully developed study would estimate the RD effect across a number of years. We felt that was beyond the scope of this study.

Table 6 RD estimates of the effect of registration in Wisconsin, 1974 and 1978

	1974	*1978*
Estimate (95% CI)	−1.5 (−5.8 to 2.6)	−2.4 (−6.6 to 1.4)

Note. Effect estimates at the threshold of municipal population of 5000. Estimates from local regression fit to both sides of the threshold with a triangular kernel. Confidence intervals based on 5000 bootstrap resamples with BC_a confidence intervals. Bandwidth selection done via mean squared error (MSE) minimization (Imbens and Kalyanaraman 2012). For the year 1974, $N=800$, and for year 1978, $N=790$.

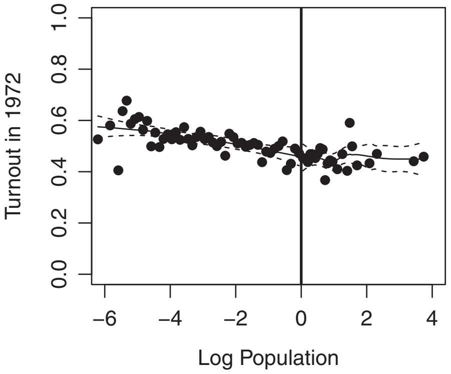

Fig. 3 RD design: the effect of voter registration system on turnout in Minnesota.

experiment departs significantly from chance. In our case, the test revealed no evidence of manipulation in either state. Thus, we have some evidence that the RD design is valid.

We first review the results for Wisconsin. In Fig. 2, for both 1974 and 1976, we plot the logarithm of population against turnout. As is standard in an RD analysis, we bin population values and plot turnout means within these bins (Lee and Lemieux 2010). We also add the fit from a nonparametric regression model and the associated 95% confidence intervals (95% CI). In both plots, we observe the same general trend, where municipalities with larger populations have lower turnout. The question, however, is whether turnout differs in a local neighborhood around the threshold of 5000. In both years, there is little evidence of a difference in turnout around this threshold.

Table 6 contains point estimates and 95% CIs for both 1974 and 1978.[19] See Appendix A for details on estimation and bandwidth selection. We first discuss the estimate from 1974, which should serve as an upper bound on the EDR effect. That is, the EDR should be as large as the registration effect or smaller. While the point estimate is in the expected direction, the point estimate is smaller than estimates from other identification strategies (−1.5 percentage points) and the standard error is too large to rule out that the effect is not zero. Even if we ignore the fact that this estimate is not statistically significant, EDR could at most have an effect of less than two percentage points.[20] The estimate for 1978, while somewhat larger, remains statistically insignificant and negative. If we were to adopt the DID assumptions, the point estimate would be in the wrong direction. Next, we examine the results from Minnesota.

Figure 3 contains a plot of turnout as a function of log population for Minnesota in 1972. Again, we bin population values and plot turnout means within these bins and add the fit from a nonparametric regression model and 95% CIs. We see that turnout tends to decline as population increases, but there is little evidence that it differs much around the threshold of 10,000.

Table 7 contains the point estimate and 95% CIs for the registration effect for Minnesota in 1972. The estimate from 1972 should serve as an upper bound on the EDR effect. That is, the effect

[19] Estimates from the fuzzy RD design were identical to those from the sharp design. As such, we only report estimates for the sharp design.

[20] Our estimate here is consistent with other work on the effect of registration, which also finds an effect of approximately two percentage points (Burden and Neiheisel 2011b).

Table 7 RD estimates of the effect of voter registration in Minnesota, 1972

	1972
Estimate (95% CI)	0.7 (−4.7 to 1.9)

Note. Effect estimates at the threshold of municipal population of 10,000. Estimates from local regression fit to both sides of the threshold with a triangular kernel. Confidence intervals based on 5000 bootstrap resamples with BC_a confidence intervals. Bandwidth selection done via MSE minimization (Imbens and Kalyanaraman 2012). $N = 815$.

EDR should be as large as or smaller than the effect of a more stringent voter registration requirement. In Minnesota, the difference is less than one percentage point and the confidence interval covers zero. If we ignore variability in the estimate, the registration effect is not even correctly signed. Thus, in both states, we find scant evidence of a registration effect, which would imply that the effect of EDR would have been negligible.

The estimate for the effect of EDR has declined from ten points under the strongest assumption to less than one or two points under the weaker assumptions of the RD design. What might explain the large difference between the estimate from the RD design and the estimates from logistic regression and DID? Is it simply that the estimands are different? We cannot provide a definitive answer, but we offer that the RD estimate differs since it creates a better counterfactual comparison in two ways. First, this is a within-state design where all the state-level confounders, of which there are many, are held constant. Second, within the states of Wisconsin and Minnesota, the inference is confined to municipalities that are actually comparable. We could use all the data, but does it make sense to compare Milwaukee or the twin cities to towns where the population is less than 500? It is possible that EDR was effective in Milwaukee or Minneapolis–St Paul, where socioeconomic status may vary more widely, but we have no good within-state counterfactual for these cities. What we can say is that for those municipalities that are comparable, there is little evidence of an EDR effect.

3 Discussion

Causal inference with observational data must invariably rely on strong untestable assumptions. In this essay, we have delineated the most commonly invoked assumptions and used them in a case study of EDR and whether it increases turnout. Techniques like matching or DID are often invoked as silver bullets, which allow one to easily estimate causal effects. But this is simply not true. DID may be very plausible in one context, but much less plausible in another. We argue that DID is less plausible in the context of turnout, since changes in the dynamics of elections from one year to the next may invalidate the key assumption. Without carefully specifying the underlying assumptions, inspecting the plausibility of those assumptions, and probing the sensitivity of inferences, it is difficult to make the move from correlation to causation. While assumptions are unavoidable in the study of politics, what needs to be clear is the role that assumptions play in the inference.

In general, we would argue that analysts should rely on a design-based inference. The design-based approach places explicit emphasis on reducing heterogeneity, clarity about identifying assumptions, a concern about endogeneity, and the role of research design (Imbens 2010, 403). The design-based approach emphasizes that without a strong research design or a credible natural experiment, complex statistical modeling cannot give correlations a causal interpretation. The concepts we presented in Sections 1.3 and 1.5 are from this design-based literature. Even with a design-based inference, much can go wrong (Caughey and Sekhon 2011). Such are the perils of trying to estimate causal effects with observational data. As we have shown, the magnitude of our statistical estimates varied widely depending on what assumptions we used. No single study, including ours, is likely to be definitive, but when the role of assumptions is transparent, the scientific community can more readily evaluate the credibility of empirical evidence. In our EDR case study, it was the design-based elements that illuminated the weakness of specification assumptions, and it was not until we used the stronger design offered by RD that our estimates became credible.

Finally, while our main goal is to present a methodological argument, we believe our study has substantive implications as well. We demonstrated that EDR appears to have done little to change turnout even in Wisconsin and Minnesota. This may explain why states that later adopted EDR have seen few signs of increased turnout. In general, this challenges much of what we know about how state institutions affect turnout.

Appendix A

A.1 Assumptions

We treated assumptions in a less formal manner in the text. Here, we present the various assumptions with formal notation using the potential outcomes framework.

A.1.1 Cross-Sectional Specification

The cross-sectional specification assumption can be written as a conditional ignorability assumption:

Assumption 1. For any unit, the potential outcomes are independent of treatment assignment once we condition on the treatment assignment mechanism:

$$Y_1,\ Y_0 \perp D \mid \mathbf{X},$$

where **X** represents a matrix of variables that confound treatment with outcomes. This assumption can be written in a variety of other ways.

A.1.2 DID

The identifying assumption for the DID estimator of treatment effects is:

Assumption 2. Conditional on the covariates, expected potential outcomes for treated and control units follow parallel paths in the absence of treatment. In formal terms,

$$E[Y_0(1) - Y_0(0)|D = 1] = E[Y_0(1) - Y_0(0)|D = 0].$$

A.1.3 Partial Identification

We consider two common assumptions to improve upon the no-assumption bounds. The first assumption that we adopt is MTR (Manski 1997). Under MTR, we assume

$$Y_1 \geq Y_0 \text{ or } Y_1 \leq Y_0. \qquad \textbf{(A1)}$$

The second assumption we consider to sharpen the inference is that of MTS. Formally, we express the MTS assumption as

$$Pr[Y_1 = 1|D = 1] \geq Pr[Y_1 = 1|D = 0]$$
$$Pr[Y_0 = 1|D = 1] \geq Pr[Y_0 = 1|D = 0].$$

A.1.4 IVs

IVs require some additional notation. The treatment indicator remains $D \in \{0,1\}$, but we introduce $Z \in \{0,1\}$ as the indicator for encouragement. Typically, Z is referred to as the instrument. In the IV setting, we seek to estimate the effect of D on Y using Z. For the IV estimand to be identified requires the following assumptions.

Assumption 3. Random assignment of the instrument:

$$Pr(Z = 1) = Pr(Z = 0).$$

Assumption 4. SUTVA (Rubin 1978):

$$\text{If } Z = Z', \text{ then } D(t) = D(t')$$

and

$$\text{If } Z = Z' \text{ and } D = D', \text{ then } Y(t,d) = Y(t',d').$$

If these two assumptions hold, one can estimate what is known as intention-to-treat effects without any further assumptions. This is simply the effect of Z on Y. To estimate the effect of D on Y requires three additional assumptions.

Assumption 5. Exclusion restriction:

$$Y(1,\ d) = Y(0,\ d) \text{ for } d = 0{,}1.$$

In words, the exclusion restriction states that the effect of Z on Y must be entirely through the effect Z_i has on D_i, or Z must not have any direct effect on Y.

Assumption 6. Nonzero average causal effect of Z on D:

$$E[D(1) - D(0)] \neq 0.$$

Assumption 7. Monotonicity (Imbens and Angrist 1994):

$$D_i(1) \leq D_i(0) \text{ for all } i = 1, \ldots, N.$$

A.1.5 RD

Hahn, Todd, and van der Klaauw (2001) demonstrate that for an RD to be identified, the potential outcomes must be a *continuous* function of the score. Under this continuity assumption, the potential outcomes can be arbitrarily correlated with the score, so that, for example, people with higher scores might have higher potential gains from treatment. This continuity assumption can be formally stated as:

Assumption 8. The conditional regression functions are continuous in s at c:

$$\lim_{s \to c} E(Y_0|S = c) = E(Y_0|S = c)$$

$$\lim_{s \to c} E(Y_1|S = c) = E(Y_1|S = c).$$

Since $Y = Y_1$, when $D = 1$, $Y = Y_0$, when $D = 0$, and $D = \mathbf{1}\{S \geq c\}$, where $\mathbf{1}\{\cdot\}$ is the indicator function, Assumption 8 implies

$$\lim_{s \to c^+} E(Y|S = c) = E(Y_1|S = c)$$

and

$$\lim_{s \to c^-} E(Y|S = c) = E(Y_0|S = c),$$

which is a formal statement of the idea that individuals very close to the cutoff, but on opposites sides of it, are comparable or good counterfactuals for each other. Thus, continuity of the conditional regression function is enough to identify the ATE *at the cutoff*. That is, the RD design identifies a LATE for the subpopulation of individuals whose value of the score is at or near c. Without further assumptions, such as constant treatment effects, the effect at c might or might not be similar to the effect at different values of S. Thus, under the continuity assumption, the RD identifies the following treatment effect:

$$\tau = E\big\{Y_1 - Y_0|(S = c\big\} = \lim_{s \to c^+} E\{Y|S = c\} - \lim_{s \to c^-} E\big\{Y|(S = c\big\}.$$

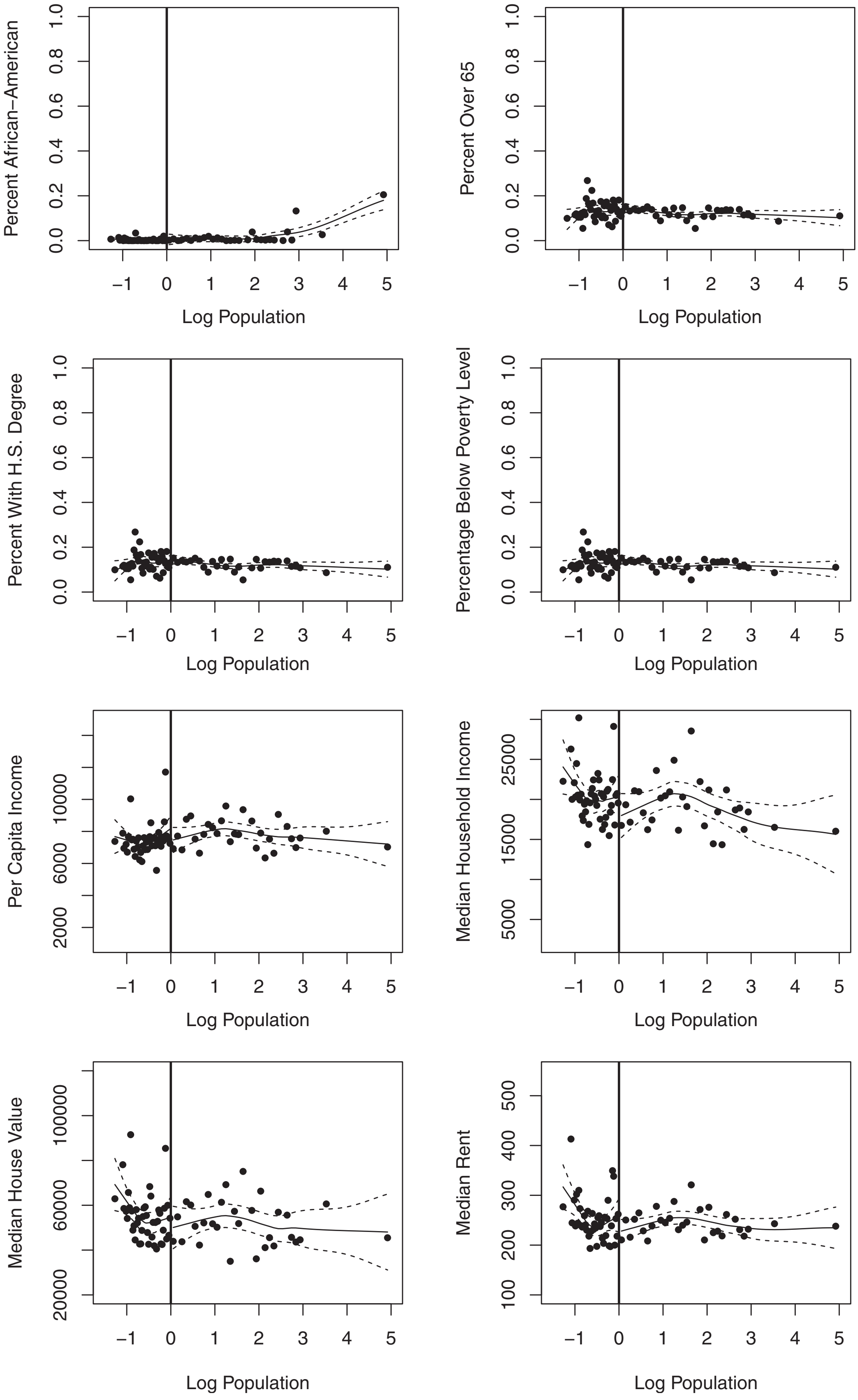

Fig. 4 Testing for discontinuities in other census covariates, Wisconsin.

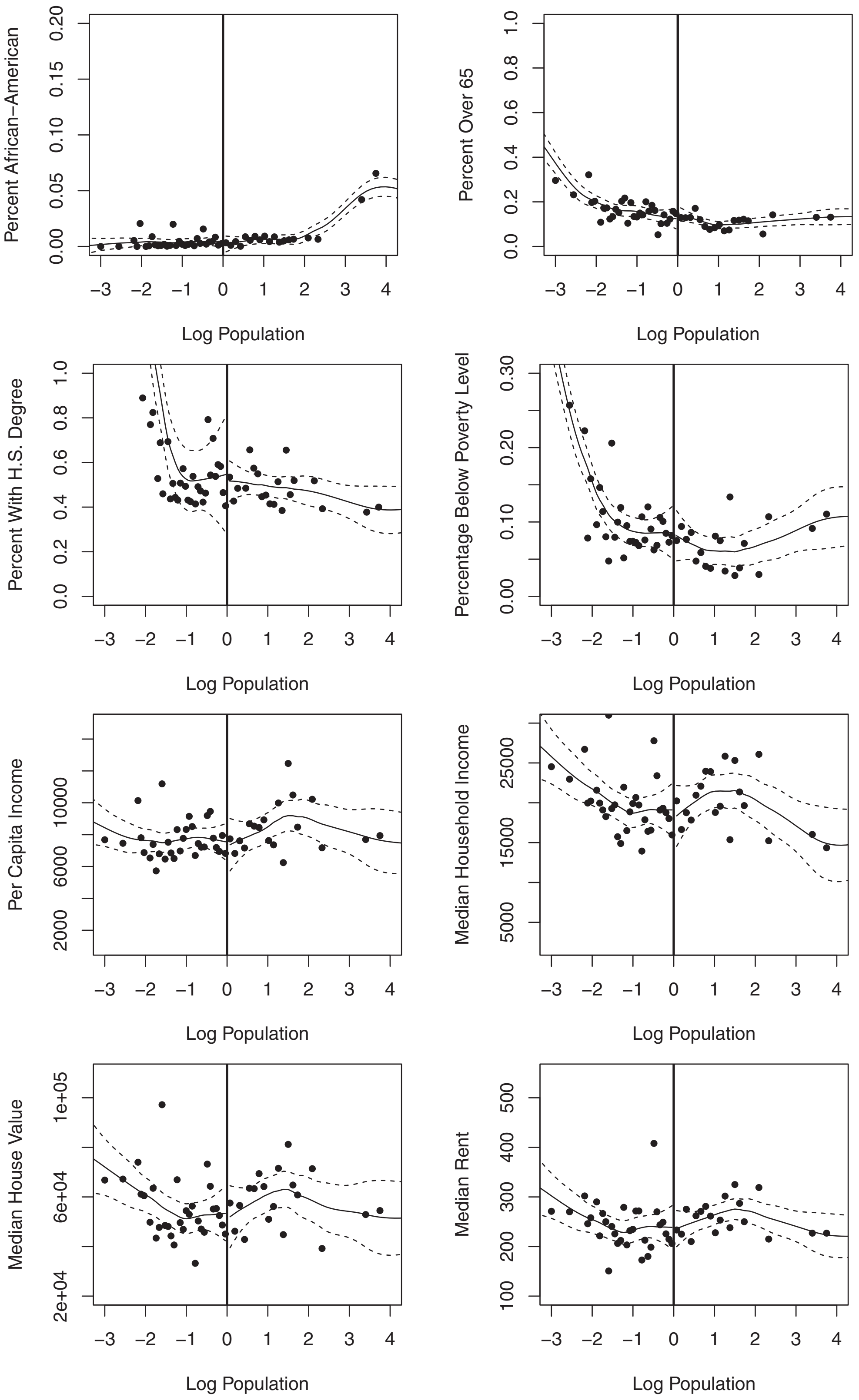

Fig. 5 Testing for discontinuities in other census covariates, Minnesota.

Estimation of this treatment effect proceeds by selecting a subset of units just above and below the discontinuity and calculating the difference across these two groups.

A.2 RD Analysis and Data

In recent years, a number of methods have been proposed for estimation of RD estimates outside simple plots. There are two related issues that analysts must contend with when estimating treatment effects in the RD design. First, one must select a local neighborhood around the discontinuity. In the RD design, we believe that observations near the cutoff are good counterfactuals; the question is how far an observation must be from the cutoff before we think observations are no longer good counterfactuals. Therefore, the analyst must select some local neighborhood above and below the cutoff. Two methods that are widely used to select the size of the local neighborhood are cross-validation (Imbens and Lemieux 2008) and algorithmic MSE minimization (Imbens and Kalyanaraman 2012).

Once the size of the local neighborhood is chosen, one can estimate either an unweighted mean difference or use (local) linear regression to estimate a conditional expectation for each side of the discontinuity and take the difference in these conditional expectations. In these methods, all observations in the local neighborhood receive equal weight. One can also use a kernel function to give observations closer to the discontinuity greater weight than those observations farther from the cutoff (Imbens and Kalyanaraman 2012). Inference can proceed either via large sample standard errors or the bootstrap. We found that no matter which method we used, the estimates and our inferences were unchanged. Our inferences were insensitive to a wide range of local neighborhood width choices. We report estimates with bandwidth selected via MSE minimization and using a triangular kernel function with local regression. For inference, we use bias-corrected and accelerated ($\mathrm{BC_a}$) bootstrap confidence intervals. We also used cross-validation to select the local neighborhood, but this made little difference.

One diagnostic that allows the reader to understand whether the score is correlated with other covariates known to be correlated with the outcome is to use these covariates as outcomes in the RD analysis. For the subset of municipalities in both states with census covariates, we plot each covariate against the score in Figs. 4 and 5. We observe no evidence of any obvious correlation between these covariates and the score at the discontinuity in either state.

The basic raw data were obtained from the Wisconsin Bluebook for the years 1975 and 1979, which contain election data from the 1974 and 1978 elections, and the Minnesota Bluebook for 1973. The Bluebook for both states contains two important sources of data. First, the Bluebook lists population based on the 1970 census by municipality. The Bluebook also records the number of votes for the two major party candidates and in each year for one third-party candidate for the gubernatorial or presidential election by municipality, and then for larger cities, the votes are further broken down by precinct. We then summed across the vote totals for the three candidates and summed by precinct for the larger cities. This provided us with a count of the number of votes for the highest office on the ballot. The Wisconsin Bluebooks are available online as PDFs, while we purchased copies of the Minnesota Bluebook and scanned them to create PDFs. We paid a data entry firm to enter the tables of population data and election returns from these PDFs. We paid for double entry, so each data table was entered separately by two different data entry specialists. We then merged these data with the municipal population data. The process, while tedious, was relatively straightforward.

References

Abadie, Alberto. 2005. Semiparametric difference-in-difference estimators. *Review of Economic Studies* 75(1):1–19.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics*. Princeton, NJ: Princeton University Press.

———. 2010. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives* 24(2):3–30.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434):444–55.

Ansolabehere, Stephen, and David M. Konisky. 2006. The introduction of voter registration and its effect on turnout. *Political Analysis* 14(1):83–100.

Barabas, Jason. 2004. How deliberation affects policy opinions. *American Political Science Review* 98(4):687–702.

Barnow, B. S., G. G. Cain, and A. S. Goldberger. 1980. Issues in the analysis of selectivity bias. In *Evaluation studies*, eds. E. Stromsdorfer and G. Farkas, Vol. 5. San Francisco: Sage.

Brians, Craig Leonard, and Bernard Grofman. 1999. When registration barriers fall, who votes? An empirical test of a rational choice model. *Public Choice* 99:161–76.

———. 2001. Election day registration's effect on U.S. voter turnout. *Social Science Quarterly* 82:170–83.

Burden, Barry C., and Jacob R. Neiheisel. 2011a. The impact of election day registration on voter turnout and election outcomes. *American Politics Research* 20(4):636–64.

———. 2011b. Election administration and the pure effect of voter registration on turnout. *Political Research Quarterly*. doi: 10.1177/1065912911430671.

Caughey, Devin, and Jasjeet S. Sekhon. 2011. Elections and the regression discontinuity design: Lessons from close U.S. house races, 1942–2008. *Political Analysis* 19(4):385–408.

Cook, T. D., and W. R. Shadish. 1994. Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology* 45:545–80.

Cook, T. D., W. R. Shadish, and Vivian C. Wong. 2008. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management* 27(4):724–50.

Donald, Stephen G., and Kevin Lang. 2007. Inference with differences-in-differences and other panel data. *Review of Economics and Statistics* 89(2):221–33.

Fisher, Ronald A. 1935. *The design of experiments*. London: Oliver and Boyd.

Green, Donald P., and Alan S. Gerber. 2002. Reclaiming the experimental tradition in political science. In *Political science: The state of the discipline*, eds. Ira Katznelson and V. Milner Helen, 805–32. New York: W. W. Nortion.

Hahn, Jinyong, Petra Todd, and Wilbert van der Klaauw. 2001. Identification and estimation of treatments effects with a regression-discontinuity design. *Econometrica* 69(1):201–9.

Hanmer, Michael J. 2007. An alternative approach to estimating who is most likely to respond to changes in registration laws. *Political Behavior* 29(1):1–30.

———. 2009. *Discount voting*. New York: Cambridge University Press.

Highton, Benjamin, and Raymond E. Wolfinger. 1998. Estimating the effects of the national voter registration act of 1993. *Political Behavior* 20(1):79–104.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81:945–60.

Imbens, Guido W. 2003. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review Papers and Proceedings* 93(2):126–32.

———. 2010. Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature* 48(2):399–423.

Imbens, Guido W., and Joshua D. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62(2):467–76.

Imbens, Guido W., and Karthik Kalyanaraman. 2012. Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies* 79(3):933–59.

Imbens, Guido W., and Thomas Lemieux 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142(2):615–35.

Just, Marion. 2011. *Should voting in the United States be mandatory?* http://www.nytimes.com/roomfordebate/2011/11/07/\\should-voting-in-the-us-be-mandatory-14/same-day-voter-registration-would-improve-turnout. (accessed December 17, 2012).

Keele, Luke J. 2012. *Replication data for: How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data*. http://hdl.handle.net/1902.1/19190, IQSS Dataverse Network [Distributor] V1 [Version]. (accessed December 17, 2012).

Knack, Stephen. 2001. Election-day registration: The second wave. *American Politics Research* 29:65–78.

Lee, David S. 2008. Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics* 142(2):675–97.

Lee, David S., and Thomas Lemieux. 2010. Regression discontinuity designs in economics. *Journal of Economic Literature* 48(2):281–355.

Manski, Charles F. 1990. Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings* 80(2):319–23.

———. 1995. *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.

———. 1997. Monotone treatment response. *Econometrica* 65(5):1311–34.

———. 2007. *Identification for prediction and decision*. Cambridge, MA: Harvard University Press.

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

Mill, John Stuart. 1867. *A system of logic: The principles of evidence and the methods of scientific investigation*. New York: Harper & Brothers.

Mitchell, Glenn E., and Christopher Wlezien. 1995. Voter registration and election laws in the United States, 1972–1992. *Inter-University Consortium for Political and Social Research* 6496:999.

Rhine, S. L. 1995. Registration reform and turnout change in American states. *American Politics Quarterly* 23:409–27.

Rosenbaum, Paul R. 2002a. Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association* 97(457):1–10.

———. 2002b. *Observational studies*. 2nd ed. New York: Springer.

———. 2005a. Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *American Statistician* 59(2):147–52.

Rosenbaum, Paul R. 2005b. Observational study. In *Encyclopedia of statistics in behavioral science*, eds. S. Everitt Brian and C. Howell David, Vol. 3, 1451–62. Hoboken, NJ: John Wiley and Sons.

Rosenbaum, Paul R. 2010. *Design of observational studies*. New York: Springer.

Rosenbaum, Paul R., and Jeffrey H. Silber. 2009. Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association* 104(488):1398–405.

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 6:688–701.

———. 1978. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* 6:34–58.

Rusk, Jerrold G. 2001. *A statistical history of the American electorate*. Washington, DC: Congresional Quarterly Press.

Sekhon, Jasjeet S. 2011. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software* 42(7):1–52.

Sekhon, Jasjeet S., and Alexis Diamond. Forthcoming. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics & Statistics*.

Sekhon, Jasjeet S., and Rocío Titiunik. 2012. When natural experiments are neither natural nor experiments. *American Political Science Review* 106(1):35–57.

Smolka, Richard G. 1977. *Election day registration: The Minnesota and Wisconsin experience in 1976*. Washington, DC: American Enterprise Institute for Public Policy Research.

Sovey, J. Allison, and Donald P. Green. 2011. Instrumental variables estimation in political science: A readers' guide. *American Journal of Political Science* 55(1):188–200.

Teixeira, Ruy A. 1992. *The disappearing American voter*. Washington DC: Brookings.

Timpone, Richard J. 1998. Structure, behavior, and voter turnout in the United States. *American Political Science Review* 92(1):145–58.

Wolfinger, Raymond E., and Steven J. Rosenstone. 1980. *Who votes?* New Haven, CT: Yale University Press.