doi:10.1017/S0003055411000542

# When Natural Experiments Are Neither Natural nor Experiments

JASJEET S. SEKHON *University of California, Berkeley*
ROCÍO TITIUNIK *University of Michigan*

*Natural experiments help to overcome some of the obstacles researchers face when making causal inferences in the social sciences. However, even when natural interventions are randomly assigned, some of the treatment–control comparisons made available by natural experiments may not be valid. We offer a framework for clarifying the issues involved, which are subtle and often overlooked. We illustrate our framework by examining four different natural experiments used in the literature. In each case, random assignment of the intervention is not sufficient to provide an unbiased estimate of the causal effect. Additional assumptions are required that are problematic. For some examples, we propose alternative research designs that avoid these conceptual difficulties.*

A natural experiment is a study in which the assignment of treatments to subjects is haphazard and possibly random. Such experiments have become increasingly prominent in recent years, and they have been used by scholars in a wide variety of fields to help make causal inferences, including political participation (Krasno and Green 2008; Lassen 2005), elections (Carman, Mitchell, and Johns 2008; Gordon and Huber 2007), political psychology (van der Brug 2001), ethnic politics (Abrajano, Nagler, and Alvarez 2005), comparative politics (Posner 2004), bureaucracy (Whitford 2002), and history (Diamond and Robinson 2010).

Natural experiments share some features with randomized experiments, but there are key differences that give rise to both inferential and conceptual problems that are often overlooked. In this type of experiment, it is often not immediately obvious which groups are comparable, leading researchers to often compare the wrong groups. Moreover, the valid comparison may not estimate the specific causal effect in which researchers are interested, but some other causal effect instead. Although these issues are critical in any research design, they are more pressing in a natural experiment where, by definition, the assignment of subjects to groups is outside the control of the researcher. As discussed later, arbitrary events or interventions are appealing to study because they may provide a useful source of variation, but they often require additional assumptions to allow for the comparisons that researchers want to make. Researchers often fail to realize that these conceptual problems exist and consequently fail to explicitly state the additional assumptions required, to use the best design for their substantive question, or to realize that they are no longer answering the causal question they began with. We propose a framework for clarifying the issues involved.

In the simplest randomized controlled experiment, subjects are randomly divided into two groups: One group is exposed to a treatment condition and the other is exposed to a control condition—usually the absence of treatment. The experiment compares the outcomes of those assigned to the treatment condition—the "treatment group"—to the outcomes of those assigned to the control condition—the "control group." Subjects are assigned to the treatment or control group based on a chance mechanism that is known (such as flipping a coin). This random assignment prevents subjects from self-selecting or being selected by others into particular groups and thus ensures that the treatment and control groups are similar in terms of all observed and unobserved characteristics. Because the only systematic difference between the two groups is the treatment that was randomly assigned, comparing the outcomes of the two groups provides an estimate of the causal effect of treatment.

Natural experiments differ from randomized controlled experiments in two fundamental ways, the first of which is commonly recognized and the second of which is not. First, the mechanism that assigns subjects to treatment and control groups is not usually known to be random. Rather, an event occurs in the world that happens to affect some subjects but not others, and the researcher assumes that the naturally occurring intervention was assigned *as-if* at random (Dunning 2008). The condition of *as-if* randomness is sometimes

Jasjeet S. Sekhon is Associate Professor, Travers Department of Political Science and Department of Statistics, University of California at Berkeley, 210 Barrows Hall, #1950, Berkeley, CA 94720 (sekhon@berkeley.edu).

Rocío Titiunik is Assistant Professor, Department of Political Science and Faculty Associate, Center for Political Studies, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48104 (titiunik@umich.edu).

A previous version of this article was circulated under the title "Exploiting Tom DeLay: A New Method for Estimating Incumbency Advantage and the Effect of Candidate Ethnicity on Turnout" and received the 2008 Robert H. Durr Award for best paper applying quantitative methods to a substantive problem during the 66th Midwest Political Science Association Annual National Conference, April 3–6, 2008. For valuable comments we thank the *APSR* editors, the anonymous reviewers, Steve Ansolabehere, Jake Bowers, Devin Caughey, Pradeep Chhibber, Daniel Enemark, Bob Erikson, Erin Hartman, John Henderson, Daniel Hidalgo, Shigeo Hirano, Francesca Jensenius, Luke Keele, Gary King, Walter Mebane, Jr., Rebecca Morton, Eric Schickler, Jonathan Wand, and participants of the Society for Political Methodology's Annual Summer Meeting 2007, 2009 Conference on Empirical Legal Studies and seminar participants at Berkeley, Columbia, Michigan, New York University, Princeton, Yale, and the University of California at San Diego. We thank Sam Davenport in the Texas Legislative Council, Nicole Boyle in the California Statewide Database, and Gary Jacobson and Jonathan Wand for providing data. Sekhon also thanks the MIT Department of Political Science for hospitality during the summers of 2007 and 2008. All errors are our responsibility.

referred to as *exogeneity*. In most studies, researchers go to great lengths to argue that this condition is satisfied. Exogeneity implies that the treatment and control groups created by the natural experiment are similar in terms of all observed and unobserved factors that may affect the outcome of interest, with the exception of the treatment and confounders that the researcher controls for. If the two groups are similar in this way, then the design is said to be valid, and comparing outcomes across the groups identifies the causal effect of treatment.

The second distinguishing feature of natural experiments is more subtle. The naturally occurring intervention generates some subjects who receive treatment and other subjects who do not. It is often possible, however, to define a number of different treatment and control groups from the natural intervention. Yet, only some of these groups are similar and thus valid to compare, even when nature intervenes randomly. The problem arises because the researcher does not directly control the design of the experiment. In a randomized controlled experiment, the researcher picks what the treatment and controls groups should be and then randomly assigns subjects to the groups. In a natural experiment, however, the researcher finds some intervention that has been implemented and also finds some subjects. She then constructs treatment and control groups to address a particular hypothesis. But the treatment and control groups constructed post hoc may not be comparable, even if one assumes that the natural intervention was randomly assigned.

In this article, we develop a framework to clarify these issues. Given a particular natural experiment, we grant the assumption that the intervention is indeed random. Once random assignment has been assumed, the task is to assess whether the treatment and control groups the researcher wishes to compare are similar. To do this, we ask two questions: (1) Is the proposed treatment–control comparison guaranteed to be valid by the assumed randomization? (2) If not, what is the comparison that is guaranteed by the randomization, and how does this comparison relate to the comparison the researcher wishes to make? We formalize these questions using the potential outcomes notation (Neyman [1923] 1990; Sekhon 2010). We also analyze several examples in different areas of political science where the answer to the first question is negative, and the answer to the second question reveals that the natural experiment in question is only indirectly related (and sometimes not at all) to the effect of interest. In these examples, additional assumptions are required that are not guaranteed to hold by the assumed randomization, and these extra assumptions are not explicitly made by the authors. In some of the examples, we offer alternative designs that make fewer assumptions.

We analyze four different natural experiments across different fields. The first example we analyze is the use of redistricting as a natural experiment to study the personal vote, a research design proposed by Ansolabehere, Snyder, and Stewart (2000) and also used by Desposato and Petrocik (2003) and Carson, Engstrom, and Roberts (2007). This is our most developed example in that we offer a new design and implement it using new data. Ansolabehere, Snyder, and Stewart (2000) exploit the fact that after redistricting most incumbents face districts that contain a combination of old and new territory, and hence they face a combination of old and new voters. The personal vote is estimated as the difference between an incumbent's vote share among new voters (voters residing in the new part of the district) and the vote share among old voters (voters residing in the old part of the district). The strength of the design is that old and new voters observe the same challenger and experience the same campaign, which means that any observed difference in incumbent vote shares between both groups cannot be attributed to differences in these factors.

Despite its desirable properties, this design faces crucial, if subtle, methodological challenges that are discovered only after a careful examination of the comparability of old voters and new voters. We show that the design would result in incorrect estimates even if voters were redistricted at random (Question 1) and that randomization guarantees the validity of a comparison that is not appropriate for estimating the personal vote (Question 2). To overcome these difficulties, we propose a new design for estimating the personal vote that uses successive implementations of multiple redistricting plans. We illustrate our design empirically using congressional elections in Texas, where two different redistricting plans were successively implemented in 2002 and 2004. We also propose a "second-best" design for cases when multiple redistricting plans are not available, which we illustrate with data from congressional elections in California and Texas.

Our second example is the impact of representatives' decisions to move from the U.S. House to the U.S. Senate on their roll-call voting scores. Grofman, Griffin, and Berry (1995) consider such moves to be a natural experiment, and they use the experiment to test different hypotheses about how changes in district preferences influence roll-call voting behavior. This example has the same logical structure, and the same problems, as our redistricting example.

Third, we consider the impact of Indian randomized electoral quotas for women on the probability that women will contest and win elections after the quotas are withdrawn (Bhavnani 2009). We demonstrate that the validity of the estimates depends on an assumption that is substantively problematic. Our reanalysis of the data shows that the evidence for the effects of the quotas is weaker than originally reported. This example illustrates that the issues we highlight can arise even when the treatment is actually randomly assigned.

For our fourth example, we consider regression discontinuity designs that have been used, among other things, to estimate the incumbency advantage in U.S. House elections (Lee 2008). In these designs, observations that lie just below or above a fixed threshold are treated as if they had been randomly assigned to be above or below the threshold. In its application to elections, candidates are considered as-if randomly assigned to winning or losing in very close elections, making bare losers and bare winners comparable. This

example is natural for us to consider given our previous discussion of incumbency and because regression discontinuity is becoming one of the more commonly used natural experimental designs. We examine an article on how time in office affects the size of the incumbency advantage (Butler 2009). The author uses regression discontinuity to estimate the difference between freshmen and nonfreshmen incumbents in their general election vote shares. We show that the study is invalid without additional assumptions, and we provide evidence that the assumptions are false.

In the next section, we examine previous uses of redistricting as a natural experiment to estimate the personal vote, propose our new research design, and implement it using data from Texas and California. In the following section, we analyze our second example about U.S. House members who move to the Senate. Next, we discuss our third example, the use of randomized gender quotas, followed by our fourth example about regression discontinuity and incumbency. We offer concluding remarks in the last section and present formal details for our examples in the Appendix.

## REDISTRICTING AS A NATURAL EXPERIMENT FOR THE PERSONAL VOTE

Ansolabehere, Snyder, and Stewart (2000) use redistricting as a natural experiment to estimate the personal vote. Redistricting induces variation in at least two dimensions: a time dimension, as voters vote both before and after redistricting, and a cross-sectional dimension, as some voters are moved to a different district while others stay in the district they originally belonged to. This natural experiment compares voters who are moved to a new district to voters whose district remains unchanged across elections, and it estimates the personal vote as the difference in the incumbent vote shares between both groups. The design is an effort to surmount the well-known methodological challenges to the estimation of the incumbency advantage (see, e.g., Cox and Katz 2002; Erikson 1971; Gelman and King 1990).

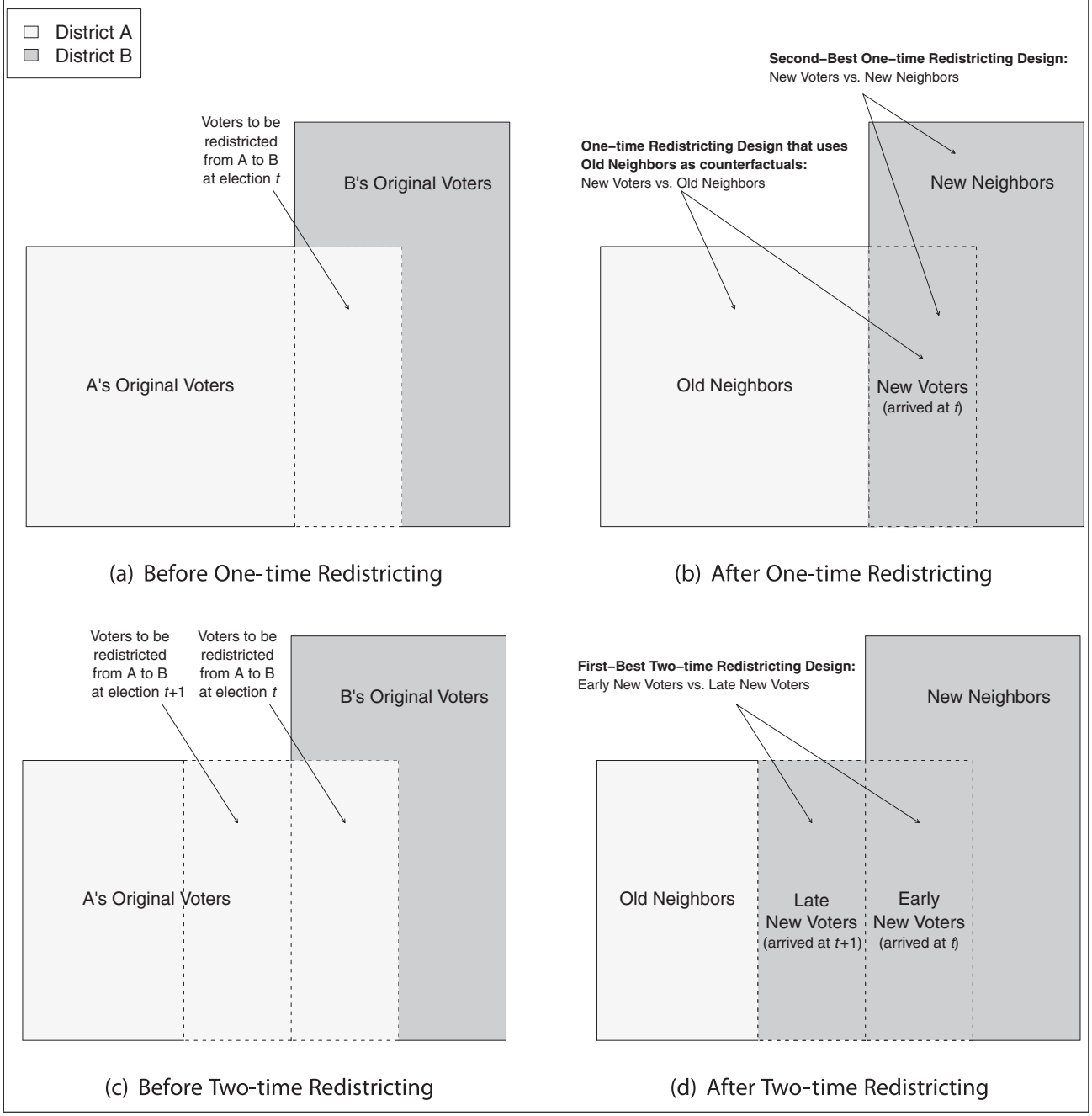
Following the approach we outlined in the introduction, we start by assuming that redistricting is truly an as-if random or exogenous intervention. We use a thought experiment to discuss the main points, which we illustrate in Figures 1(a) and 1(b) and derive formally in the Appendix. We imagine that just before election $t$ a redistricting plan randomly redraws the boundaries of district A, in such a way that some voters are randomly chosen from district A and moved to district B. In the first post-redistricting election, the incumbent in district B now represents some new voters who previously resided in district A and some old voters who remain in district B. It seems natural to attribute the differences in how these two groups vote to the personal incumbency advantage of the incumbent, as in Ansolabehere, Snyder, and Stewart (2000). After redistricting, both groups of voters face the same incumbent, challenger, and general campaign environment, but the groups differ in their history with the incumbent.[1]

We first consider Question (1): Is the comparison of B's old voters and B's new voters guaranteed to be valid if we assume that voters are redistricted randomly? It is simple to show that it is not. Although this randomization guarantees that voters who stay in A and voters who leave A are comparable, randomization says nothing about the comparability of B's new voters and B's old voters. This is clearly illustrated in Figures 1(a) and 1(b), where we refer to voters who are in district A before and after redistricting as *Old Neighbors*, voters who are in A before redistricting and in B after redistricting as *New Voters*, and voters who are in B before and after redistricting as *New Neighbors*. Because New Voters and New Neighbors are in a different district before redistricting, they will have different histories by the time redistricting is implemented. For example, New Voters may have been moved from a Hispanic to a non-Hispanic incumbent or from a Democratic to a Republican incumbent, whereas New Neighbors will face no variation in the characteristics of their incumbent (assuming their incumbent does not retire after redistricting). Different previous histories will likely affect how New Voters react to their new incumbent, thus observed differences in incumbent vote shares among these groups cannot solely be attributed to the fact that B's incumbent is known to New Neighbors (who have always been in B) but unknown to New Voters (who arrived in B after redistricting). Random assignment does not make the previous history of both groups comparable; if we wish to use New Voters and New Neighbors to learn about the personal vote, such comparability must be assumed *in addition* to the randomization.

The methodological design is complicated by an ambiguity in the way comparison groups are defined. New Voters are naturally understood to be voters whose district changes randomly between one election and another (i.e., those who arrive at $t$ to district B, as shown in Figure 1(b)). However, we can define at least two potential comparison groups: New Neighbors and Old Neighbors. New Neighbors are the electorate of the district to which New Voters are moved, and Old Neighbors are the electorate of the district to which New Voters belonged before redistricting. In this case, the natural experiment creates three distinct groups, and not all groups are valid to compare. In particular, comparing New Voters and New Neighbors is not valid, as shown in Figure 1. Although redistricted voters are chosen randomly from district A, none of the voters in district B can participate in this randomization, and hence they are not guaranteed to be comparable to district A's voters. Henceforth, we refer to the design that compares New Neighbors and New Voters as the "second-best one-time" (SBOT) redistricting design.

[1] Although this thought experiment places constraints on which precincts may move, the results are general. That is, the conclusions are the same if we assume that every precinct in every district in the state has a positive probability of moving to any other district. However, the notation and discussion become unwieldy.

**FIGURE 1. Illustration of Redistricting Research Designs**




This leads naturally to Question (2): What is the comparison that is guaranteed to be valid if we assume that redistricting was done at random? And how does this comparison relate to the comparison between New Voters and New Neighbors proposed by Ansolabehere, Snyder, and Stewart (2000) and to the personal vote? As the earlier discussion makes clear, in this case the valid comparison is the one that involves the two groups that directly participated in the randomization. In Figure 1(a), it is clearly shown that only A's original voters were subject to the randomization. Thus, it follows that the groups that random redistricting guarantees to be valid to compare are Old Neighbors and New Voters, *both of which were in district A before redistricting*.

A natural alternative design to the SBOT design is one that compares Old Neighbors and New Voters. Given a single redistricting intervention, this comparison is the best available option because it makes the comparison that is guaranteed to be valid if the natural experiment were truly random. The remaining question is whether this comparison can be used to estimate the personal vote. Unfortunately, this design introduces important sources of heterogeneity. It compares voters

who, before redistricting, are in the same district (and hence face the same electoral environment) but who, after redistricting, are in different districts (and hence face a different incumbent, a different challenger, a different campaign, etc.). Thus, a comparison of incumbent vote shares between these two groups after redistricting will be affected not only by the change in incumbent but also by all the district-level factors that affect voting. This problem does not arise when New Voters and New Neighbors are compared, because both groups are in the same district after redistricting, an attractive feature that was at the heart of Ansolabehere, Snyder, and Stewart's (2000) original motivation for the design. Paradoxically, when a single redistricting plan is available, fixing the design to use the comparison that is made valid by the randomization requires abandoning the most desirable feature of the original design (both groups facing the same incumbent at $t$) and there by drastically changing the effect that is being estimated.

At first glance, it might seem that we could modify the comparison between Old Voters and New Voters by narrowing the set of movements between districts to include only homogeneous changes and hence increasing their comparability after redistricting. For example, we could restrict the analysis to Old Voters and New Voters who, despite being in different districts at $t$, face incumbents of the same party and challengers of equivalent quality. Unfortunately, a crucial difficulty with this approach is that, to induce the desired homogeneity, it restricts the analysis based on characteristics of the environment after redistricting. And because these characteristics are likely to have been affected by redistricting itself, we run the risk of introducing bias in the results. We therefore conclude that the one-time redistricting design that compares Old Voters and New Voters, although valid under randomization, is not appropriate to estimate the effect of incumbency status on electoral outcomes.[2]

## Consecutive Redistricting: The First-Best Design

We propose a different design. We consider a modification of the thought experiment introduced earlier, and imagine that, after some voters are randomly moved from district A to B (and after election $t$ takes place), another random redistricting plan is implemented right before election $t+1$ so that some voters who were in district A until after election $t$ are randomly chosen and moved to district B. At $t+1$, there are three types of voters in district B: voters who always belonged to B, voters who were moved to district B just before election $t$ (henceforth *Early New Voters*), and voters who were moved to district B just before election $t+1$ (henceforth *Late New Voters*). The design is illustrated in Figures 1(c) and 1(d). In this case, the most natural way to estimate the causal effect of incumbency is to compare Early New Voters to Late New Voters. No only do these two groups face the same electoral environment at election $t+1$ but they also have the same electoral environment up to election $t-1$. This feature implies that their histories are the same except for the fact that Early New Voters are moved to the new district one election earlier than Late New Voters. We call this the "first-best two-time" (FBTT) redistricting design, and show that it is free from the complications that arise in the two alternatives considered earlier.

Importantly, if voters are redistricted randomly, Early and Late New Voters will be comparable before the first redistricting plan takes effect. To make sure that both groups are still comparable just before the second redistricting plan is implemented, we need an additional assumption, namely, that in the absence of redistricting, Early New Voters and Late New Voters would have had the same trend in incumbent vote shares between election $t-1$ and election $t+1$. We provide a precise formalization of this assumption in the Appendix.

Although methodologically rigorous, our FBTT design may be somewhat limited in application. When two redistricting plans are implemented in consecutive elections, Early New Voters spend only one additional election in the new district when compared to Late New Voters, and the design can capture the incumbency advantage that accrues in a single election cycle, but not in longer periods. If we believe that some fraction of the personal vote accrues over longer periods of time, this will not be included in the effect estimated by this design. Strictly speaking, however, our design could be applied to two redistricting plans that occur, for example, 10 years apart, because nothing in its derivation requires that the interval between plans be of a specific length. The problem with longer time intervals is one of practical implementation, because in the lapse of 10 years a large proportion of voters will move to a different neighborhood, city, or state, severely limiting the comparability of voting units such as precincts or blocks over time.

Our estimand is similar to the one originally proposed by Ansolabehere, Snyder, and Stewart (2000). In their main design, these authors compared incumbent vote shares in the new and old parts of the district (New Voters and New Neighbors in our terminology) only in the first election after redistricting. Thus, both treatment groups—Late New Voters in our FBTT design and New Voters in their design—have never been represented by the new incumbent before redistricting. The only difference between the designs arises in the control group: In our case, Early New Voters have been with the new incumbent for exactly one election, whereas in their design New Neighbors (the voters in the old part of the district) have been with the incumbent an unspecified amount of time that depends on how many elections before redistricting the incumbent was first elected. Note that if one wants to suggest that in those districts where the incumbent was first elected,

[2] We note, however, that this design could be used to estimate how voters react to a change in the race or ethnicity of their incumbent, because in this case one wishes to consider the different electoral environments brought about by incumbents of different races or ethnicities.

say, 10 years before redistricting, the New Neighbors–New Voters comparison in Ansolabehere, Snyder, and Stewart (2000) measures the personal vote that accrues in 10 years, but one must again make the assumption mentioned earlier that the composition of voting units (counties in this case) is constant over this period. Just as it happened in our design, the comparability of voting units becomes more implausible the longer the time period considered. Thus, both designs require similar assumptions to be able to estimate the personal vote over long periods of time. In addition, as shown in table 3 in Ansolabehere, Snyder, and Stewart (2000), the largest part of the incumbency effect accrues by the first election after redistricting.[3] This finding suggests that any research design such as ours or Ansolabehere, Snyder and Stewart's (2000) that focuses on the first election after redistricting will likely capture the most significant part of the overall incumbency effect.

## Empirical Application of Redistricting Designs: California and Texas

We illustrate the FBTT design using data on congressional elections from Texas and the SBOT design using data on congressional elections from both Texas and California, focusing on the personal vote of U.S. House members between 1998 and 2006. Our choice of Texas is motivated by the availability of data at the Voting Tabulation District (VTD) level, which allows us to track the same geographical unit over time, and by the consecutive congressional redistricting plans implemented in 2002 and 2004, which give us the unique opportunity of implementing the FBTT design. Our choice of California, where a single redistricting plan was implemented in 2002, is motivated by the availability of data at the census block level, which, as in Texas, allows us to track the same geographical unit over time. Further details about the data and redistricting plans are provided in the supplemental Online Appendix (available at http://www.journals.cambridge.org/psr2012002).[4]

***As-if Randomness.*** So far, our methodological discussion assumed that redistricting was as-if randomly assigned. This assumption was made to illustrate the challenges that natural experiments can pose even in the best-case scenario of random assignment. Although we continue to make this assumption in the discussion of the examples that follow, in this empirical application we do examine whether as-if randomness is a plausible assumption. We believe that a thorough analysis of this assumption ought to be the first step in the empirical study of all natural experiments. We discuss this issue succinctly here and refer readers to the supplemental Online Appendix for a more detailed discussion.

We must establish whether voters who are moved to a new district are comparable to voters who are left behind. In other words, is the decision to move some voters from one incumbent to another as-if random? Or are redistricting maps drawn in such a way as to move some particular types of voters and not others? We illustrate our answer in Figure 2, which shows the empirical Quantile-Quantile (QQ) plots of the baseline vote share received by the incumbent U.S. House member in the election before redistricting, comparing units that were to be redistricted to a different incumbent in the following election (*would-be treatments*) to units that were to remain with the same incumbent after redistricting (*would-be controls*).[5] Figure 2(a) shows the QQ plot for California, whereas Figure 2(b) shows the QQ plot for Texas. Because the outcome examined is the incumbent vote share before redistricting occurs, the true effect is zero by construction. However, in both states, the empirical quantiles of the baseline incumbent vote share for would-be treatments are everywhere larger than the empirical quantiles for would-be controls, indicating that units with a lower incumbent vote share in the election before redistricting are more likely to be moved to a different incumbent when redistricting is implemented.

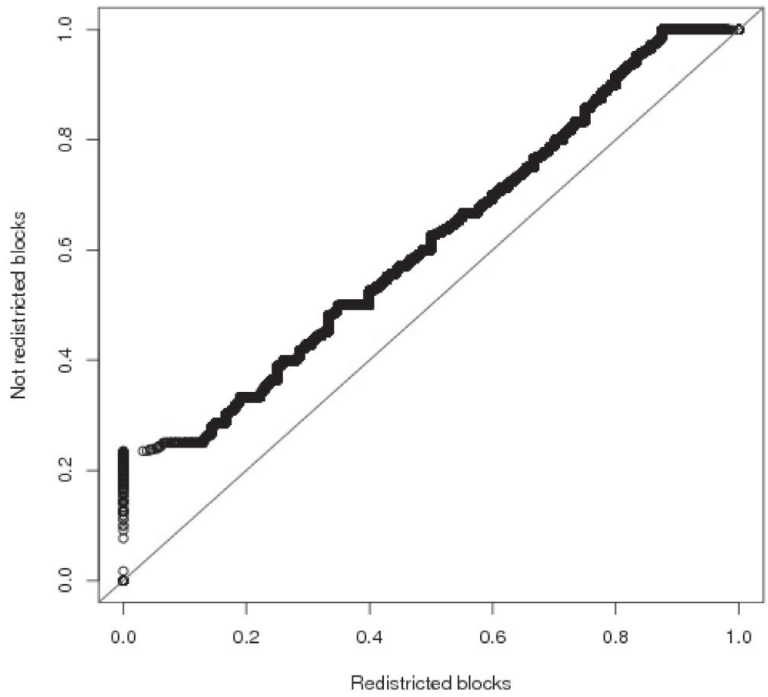
If at least part of this tendency of new voters to vote for their incumbent at a lower rate persists in the future, comparing voters who are moved to a new district to voters who are not will be biased toward finding a positive personal vote even when there is none. Thus, the evidence in Texas and California shows that redistricting, as it is, cannot be considered as-if random. However, we are able to find a group of covariates such that, after controlling for them, redistricting can be credibly shown to be exogenous. We find such covariates by means of a "placebo" test, a test in which we know that the true effect of redistricting is zero, and we look for the covariates that allow us to recover this true effect. This placebo test is based on the FBTT design in Texas, and examines VTDs that will be redistricted (or not) in 2004, but are in the same district in elections 1998, 2000, and 2002. We call those to be redistricted in election 2004 "treated" and those who will not be redistricted in this election "controls," and arbitrarily denote 2000 to be the baseline year. Our placebo test is that in 2002 there should be no significant difference

[3] This table shows the difference in incumbent vote shares between voters new and old to the district in the first, second, third, and fourth election after redistricting for the subset of incumbents who do not retire and are not defeated. In the modern period, the highest effect is seen in the first election after redistricting, with the effect decreasing monotonically over time. The personal vote effect in the first election after redistricting is about twice the effect observed in the second election after redistricting, and almost four times the effect observed in the third and fourth post-redistricting elections.

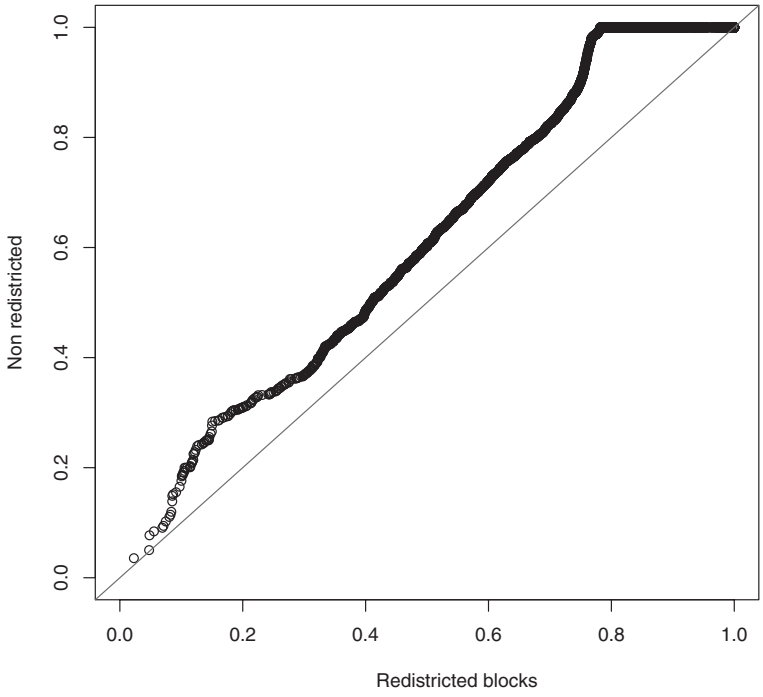
[4] It is important to note that the time period, geographical regions, and units of observation that we use in this empirical illustration are generally different from those employed in previous uses of redistricting to estimate the personal vote. For this reason, any differences (or lack thereof) between our results and those available in previous applications of redistricting cannot be directly attributed to the methodological differences between our research design and the original design proposed by Ansolabehere, Snyder, and Stewart (2000).

[5] The unit of analysis is the Voting Tabulation District for Texas, and the 2000 census block for California. Details are provided in the supplemental Online Appendix.

**FIGURE 2. QQ Plots of Baseline Vote Share for Incumbent House Member, California and Texas**




(a) California, 2000




(b) Texas, 2002

**TABLE 1. Balance for Placebo Test Covariates for Texas**

| | Before Matching | | | After Matching | | |
|---|---|---|---|---|---|---|
| Variable | Mean diff | D-statistic | KS p-value | Mean diff | D-statistic | KS p-value |
| Dem Pres. vote share '00 | .0447 | .100 | 0.00 | .00459 | .0337 | 0.953 |
| Dem House vote share '00 | .159 | .305 | 0.00 | .00693 | .0344 | 0.678 |
| Dem House vote share '98 | .127 | .340 | 0.00 | .00585 | .0368 | 0.996 |
| Dem Senate vote share '00 | .0426 | .120 | 0.00 | .00576 | .0317 | 0.846 |
| Dem Governor vote share '98 | .0305 | .0974 | 0.00 | .00510 | .0241 | 0.942 |
| Dem Att. Gen. vote share '98 | .0353 | .141 | 0.00 | .00683 | .0358 | 0.868 |
| Dem Comptroller vote share '98 | .0304 | .208 | 0.00 | .00499 | .0373 | 0.994 |
| Voter turnout '00 | .0331 | .102 | 0.00 | .00607 | .0327 | 0.943 |
| Voter turnout '98 | .028 | .199 | 0.00 | .0111 | .0378 | 0.235 |
| Registration '00 | .0308 | .157 | 0.00 | .00736 | .0608 | 0.601 |

*Notes*: The mean differences are the simple differences between treatment and control, the D-statistic is the largest difference in the empirical QQ-plot on the scale of the variable, and the KS p-value is from the bootstrapped Kolmogorov-Smirnov test.

between the incumbent vote shares of our treated and control groups.[6]

Table 1 provides the covariates we controled for in our placebo test. We ensured that these covariates were similar between treatment and control groups using Genetic Matching (GenMatch), a matching method that maximizes the similarity in observed covariates between treated and control groups (Diamond and Sekhon 2005; 2011; Sekhon and Grieve n.d.). This method creates similar pairs; that is, for every unit in the treatment group, it finds the most similar unit in the control group in terms of the observed covariates. The treatment effect is then estimated as the difference in means in the outcome of interest across these matched pairs. Table 1 shows that the treatment and control groups used in this placebo test looked very different before controlling for the covariates, but after the matching procedure they look indistinguishable in all the covariates described earlier. The mean differences between treatment and control groups, the maximum differences in the empirical QQ-plots, and the statistical significance of the differences greatly shrank post-matching in every case.

Table 2 presents the placebo results for the 2002 incumbent vote shares. In this table, as well as in Tables 3 and 4, the estimated effect shown is the difference in the average vote share between treatment and control groups across matched pairs obtained via Genetic Matching; confidence intervals are also reported.[7] As shown in Table 2, the estimate of the placebo test

**TABLE 2. Placebo Test for 2002 Incumbent Vote Share in Texas**

| | Estimate | 95% CI | | *p*-Value |
|---|---|---|---|---|
| Incumbent vote '02 | 0.00245 | −0.00488 | 0.00954 | 0.513 |

*Notes*: Estimate is difference in means in vote proportions after matching on covariates described in Table 1 using Genetic Matching. Confidence intervals are calculated using Hodges-Lehmann interval estimation. There are 474 observations.

is statistically indistinguishable from zero and substantively small (0.00245). This is not a case of the confidence interval simply being large: The point estimate is extremely small and the confidence interval tight. Thus, this placebo test shows that when we control for a rich set of covariates, there is no significant difference in vote shares for the 2002 House incumbent between VTDs that will be redistricted in 2004 and VTDs that will not, which indicates that this set of covariates is enough to recover the true zero effect. Given these results, in our analyses in the following sections, we compare the outcomes of treatment and control groups after controlling for the observable characteristics described in Table 1 via matching, just as we did in our placebo test.

***Difference Between Old Voters and New Voters When Party Remains the Same.*** The results for Texas are displayed in Table 3. Rows (1) and (2) present the results from the FBTT design when the party of the incumbent remains unchanged before and after redistricting. The difference in the vote shares of Late New Voters and Early New Voters is estimated to be statistically indistinguishable from zero for both 2004 and 2006. Note that our point estimates are also extremely small. For example, for 2004, the estimated vote proportion is 0.00637 and for 2006 it is 0.00843.

[6] A crucial feature of this placebo test is that treated and controls are always in the same district when analyzed, which implies that it can be used to validate the FBTT design but not the SBOT design, because in the FBTT design treatment and control precincts are in the same district before redistricting, whereas in the SBOT design precincts are in the same district after redistricting but in a different district before redistricting. Nonetheless, this placebo test can provide indirect evidence about the observable characteristics that should be controlled for in the SBOT design.

[7] Confidence intervals are obtained from Hodges-Lehmann Interval Estimation. See the supplemental Online Appendix for additional details.

**TABLE 3. Results for Texas**

| | | Estimate | 95% CI | | *p*-Value |
|---|---|---|---|---|---|
| FBTT design, same-party movements | | | | | |
| (1) | Incumbent vote '04 | 0.00637 | −0.00428 | 0.0177 | 0.254 |
| (2) | Incumbent vote '06 | 0.00843 | −0.00938 | 0.0258 | 0.457 |
| SBOT design, same-party movements | | | | | |
| (3) | Incumbent vote '04 | 0.00214 | −0.00807 | 0.0124 | 0.690 |
| (4) | Incumbent vote '06 | 0.00472 | −0.00539 | 0.0149 | 0.378 |
| FBTT design, different-party movements | | | | | |
| (5) | Incumbent vote '04 | 0.119 | 0.0595 | 0.191 | 0.0000 |
| (6) | Incumbent vote '06 | 0.0389 | 0.00973 | 0.0692 | 0.0106 |

*Notes*: Estimate is difference in means in vote proportions after matching on covariates described in section *Redistricting as a Natural Experiment for the Personal Vote*, using Genetic Matching. Confidence intervals are calculated using Hodges-Lehmann interval estimation. For the same-party SBOT design, there are 434 observations. For the FBTT design, there are 166 observations in the same-party case and 70 observations in the different-party case.

**TABLE 4. Results for California**

| | | Estimate | 95% CI | | *p*-Value |
|---|---|---|---|---|---|
| SBOT design, same-party movements | | | | | |
| (1) | Incumbent vote '02 | 0.0219 | 0.0171 | 0.0266 | 0.000 |
| (2) | Incumbent vote '04 | 0.0240 | 0.0195 | 0.0284 | 0.000 |
| (3) | Incumbent vote '06 | −0.0072 | −0.0129 | −0.0015 | 0.012 |
| SBOT design, different-party movements | | | | | |
| (4) | Incumbent vote '02 | 0.1025 | 0.0925 | 0.1122 | 0.000 |
| (5) | Incumbent vote '04 | 0.1020 | 0.0932 | 0.1109 | 0.000 |
| (6) | Incumbent vote '06 | 0.0307 | 0.0186 | 0.0428 | 0.000 |

*Notes*: Estimate is difference in means in vote proportions after matching on covariates described in section *Redistricting as a Natural Experiment for the Personal Vote*, using Genetic Matching. Confidence intervals are calculated using Hodges-Lehmann interval estimation. There are 3,526 observations for the same-party SBOT design and 1,394 observations for the different-party SBOT design.

The third and fourth rows in Table 3 present our estimates from the SBOT design controlling for the same covariates we used in the FBTT design (those in Table 1), with the addition of variables that attempt to measure details of the House election at baseline in 2000 and in 1998. In particular, we add Jacobson's challenger quality measures in 2000 and in 1998. As is made clear in rows (1) through (4) of Table 3, when controlling for a large set of covariates, all of our estimates of the difference in the incumbent vote shares between New Neighbors and New Voters in Texas when the party of the incumbent is unchanged are extremely small, and all are *substantively* and *statistically* indistinguishable from zero. The largest absolute value of the point estimate is 0.00472 (incumbent vote in '06). This is insignificant, but even if it were significant, it would not be a substantively meaningful effect.

The results for same-party movements for California are displayed in Table 4. As mentioned earlier, because in this state there is only one redistricting plan implemented during the 2000 decade, we cannot apply the FBTT design nor is a placebo test available to directly validate the design. Thus, we can only use the SBOT design in California. Rows (1) through (3) in Table 4 show the same-party results for the SBOT design, again controlling for a large set of covariates. Contrary to the results for Texas, in California there is a statistically significant difference between New Voters and New Neighbors as estimated by this design. In 2002, the difference in the incumbent vote shares between these two groups is 2.2%. The results for 2004, the second election after redistricting, are similar to the 2002 results in magnitude and significance. But in the 2006 election, the initial electoral advantage enjoyed by the incumbent among New Neighbors decreases. The effect appears to switch signs, but this is not a robust result. As shown in row (3) of Table 4, the effect is substantively small, just −0.72%, and the effect is not significant when alternative statistical tests are used (not shown).

***Difference Between Old Voters and New Voters When Party Changes.*** Rows (5) and (6) in Table 3 present the results for Texas from our preferred FBTT design,

but when the party of the incumbent changes after redistricting. Here we find that there is a significant and large difference in the incumbent vote shares of Early and Late New Voters. Early New Voter VTDs vote for the incumbent party at a much higher rate than Late New Voter VTDs: 11.9%. But by the time that Late New Voters have been in the congressional district for a term, this effect drops to about 3.9%.

The California results for movements to an incumbent of the opposite party using the SBOT design are shown in rows (4) through (6) of Table 4. The incumbent vote share among New Neighbors is larger than among New Voters, and the difference is large and statistically significant. In 2002, this difference is about 10%. The figure for 2004 is similar, although the effect decreases to about 3% by 2006, two elections after redistricting. But these results must be interpreted with caution. As mentioned earlier, the lack of multiple redistricting plans in California forces us to consider only the SBOT design, which becomes particularly problematic when considering movements between incumbents of the opposite party. The reason is that when movement occurs across parties, we cannot control for previous incumbent vote share because this would entail comparing units with equal vote shares for *opposite* parties, and voting for a Democratic incumbent is not comparable to voting for a Republican incumbent. The FBTT design does not suffer from this issue, because all units are in the same district before redistricting.

## LEGISLATORS' CAREER DECISIONS

Grofman, Griffin, and Berry (1995) study the voting behavior of U.S. House members who move to the U.S. Senate. In moving from the House to the Senate, legislators will be moving, on average, to a more heterogeneous constituency. Applying a Downsian logic, the authors hypothesize that, to increase their (re)election potential, Democratic members of the House will move to the right upon entering the Senate, and Republican members will move to the left (at least relative to their party's medians).

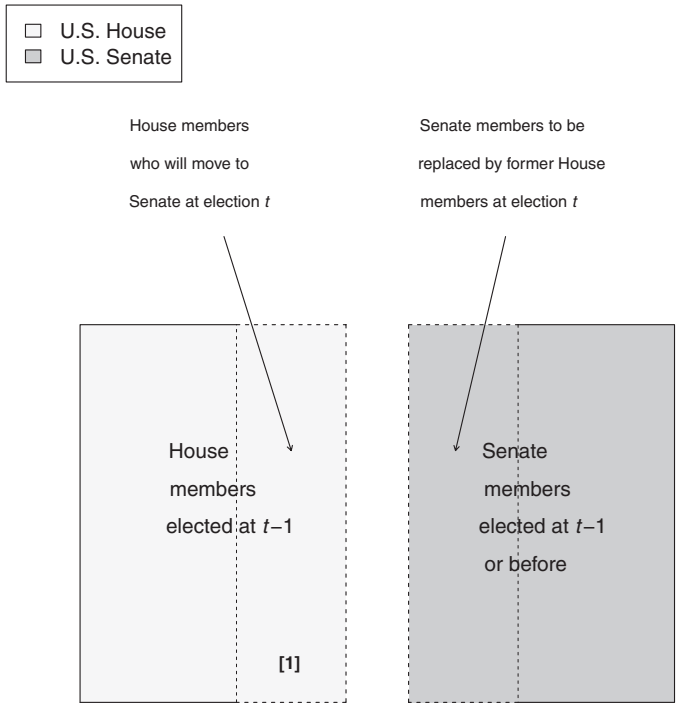
To test this hypothesis, they treat individual members' decisions to move to the Senate as a natural experiment, and they define the treatment group to be the members who move. Interestingly, the authors recognize that this movement creates several groups that can be compared with the treatment group. They define three different control groups and compare each in turn with the treatment group. These control groups are the House members left behind by treated members, the serving members of the Senate whom treated members join after leaving the House, and the treated members themselves before they move. We now use our framework to analyze the validity of these comparisons.

We assume that the decision to move to the Senate is randomly assigned among House members to consider this natural experiment in its best-case scenario: In election $t$, a random mechanism assigns some House members to move to the Senate and the rest to stay in the House.[8] Figure 3(a) illustrates the situation after election $t-1$ but before election $t$, prior to movement from House to Senate. In election $t-1$, all members of the House and a third of the members of the Senate are elected or reelected. Some of the House members elected at $t-1$ are randomly selected to win a Senate seat in the following election, at $t$. In turn, some members of the Senate elected before $t-1$ will be replaced by these incoming former House members at $t$, either because they will retire or because former House members will defeat them in the general or primary election. As mentioned earlier, the authors consider three different control groups that lead to three different research designs, which we call *Design 1*, *Design 2* and *Design 3* and illustrate in Figure 3(b). In Design 1, former House members who move to the Senate are compared to the House members they leave behind. Randomly choosing House members just before $t$ ensures that those left behind and those who move to the Senate are comparable in their observable and unobservable characteristics. This is because, as can be seen in Figure 3(a), all House members are part of the same population for which the randomization is performed before election $t$. Thus, a comparison of House members who move and House members who stay behind is valid under randomization. Note that this is exactly analogous to the comparison between New Voters and Old Neighbors in our previous example.

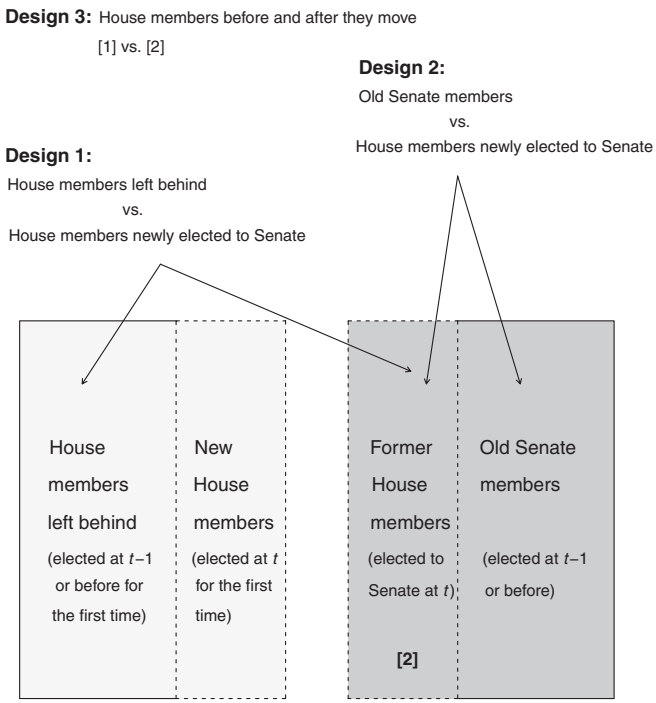
The problem with Design 1 is that, after House members move to the Senate, they face different bills, committees, party leaders, etc., from those faced by the House members they left behind. Evaluating whether they become more moderate than this control group will be complicated by the lack of comparability in their respective legislative environments. Once again, these difficulties are analogous to the difficulties that arose when Old Neighbors were compared to New Voters in the redistricting example. Recall that Old Neighbors and New Voters are in a different district after redistricting and face a different incumbent, challenger, and campaign. Similarly, House members who move to the Senate and House members left behind belong to different legislative chambers and consequently face different legislative environments. Thus, it will be difficult to attribute the difference between the groups in Design 1 solely to a change in the characteristics of the representatives' constituencies. The authors are aware of this difficulty, and perhaps for this reason they do not compare these groups after the treatment takes place, only before.

[8] In this best-case scenario, we ignore the distinction between disputing and winning a seat. If the hypothetical random experiment were only to randomly select House members to dispute a Senate seat, it would not guarantee comparability unless election outcomes were also random (or there simply were no elections). Otherwise, it seems natural to assume that better candidates would be more likely to win and that these candidates would not be comparable to the House members randomly chosen to remain in the House. We ignore this additional complication, but the difficulties in designing a thought experiment in which the treatment of interest is randomly assigned are an indication of the challenges faced by this natural experiment.

**FIGURE 3. Illustration of House-to-Senate Movements Research Designs**




(a) U.S. House and Senate after election $t-1$ and before election $t$




(b) U.S. House and Senate after election $t$

Grofman, Griffin, and Berry (1995) also compare House members who move to the "old" members of the Senate whom they join after they move. We call this Design 2. This is analogous to the comparison between New Voters and New Neighbors in the personal vote example, and it faces the same methodological challenges: Old members of the Senate are not part of the population for which the treatment was assigned. Therefore, the comparison between House members who move and their new colleagues in the Senate is not valid under random assignment alone. This can be seen by comparing Figures 3(a) and 3(b). In Figure 3(a), it is clear that before election $t$, House members who will move to the Senate and Senate members elected in or before election $t-1$ are not in the same chamber before $t$. This is by construction: House members who will move to the Senate are in the House at $t-1$, and Senate members who will be joined by former House members at $t$ are in the Senate at $t-1$. Thus, those House members who are randomly moved to the Senate may or may not have similar characteristics to the Senate members they join at $t$. If this comparison is to be informative of the moderating effects of moving from the House to the Senate, the comparability of these groups must be *assumed*.

Finally, Grofman, Griffin, and Berry (1995) make a before-and-after comparison within House members who move to the Senate. This is what we call Design 3, which simply compares House members who will move to the Senate before they move, to the same House members after they move. This over-time comparison circumvents some of the issues mentioned previously. In this design it is no longer necessary to assume that House members who move are comparable to either House members left behind or Senate members they join, because neither group is used as a comparison group. Comparing the same group of House members before and after they move implies that, by construction, many of their observable and unobservable characteristics will be exactly the same before and after, because they are the same group of representatives. However, there are two difficulties with this approach. The first one is that a change in behavior of moving House members between $t-1$ and $t$ can only be attributed to the movement from the House to the Senate if other factors that affect roll-call voting are held constant between these two periods. This may not be a plausible assumption if factors such as partisan tides, differences between on-year and off-year elections, etc., both change over time and affect representatives' legislative behavior. In other words, this over-time comparison is only meaningful under the assumption that moving House members would not have changed their behavior between $t-1$ and $t$ if they had stayed in the House. This stability assumption is not guaranteed by randomization.

The second difficulty arises by construction and is similar to the difficulties faced by Design 1. House members who move are in the House before they move, but they are in the Senate after they move. This movement is precisely what the natural experiment seeks to leverage to answer the question of interest. But, once again, this movement, by its very definition, implies that the legislative environment faced by House members before they move is different from the environment faced after they are elected to the Senate. As a result, it is difficult to attribute observed changes in legislative behavior to any particular factor. For example, if we observe that after moving to the Senate, members' roll-call voting scores become more moderate, it will be difficult to know whether this effect is attributable to a change in constituency preferences, in leadership, or in the kinds of bills that reach the Senate floor. We could make additional assumptions about how these factors change over time and how they affect legislative behavior, but these assumptions are not guaranteed to hold by the assumed random assignment to move from the House to the Senate.

## RANDOMIZED GENDER QUOTAS IN INDIA

Bhavnani (2009) studies whether electoral quotas for women alter women's chances of contesting and winning elections after these quotas are withdrawn. Few dispute that electoral quotas increase women's representation while they are in place. The open theoretical question is what happens after they are removed. To answer this question, the author uses an experiment that occurs in India in which a third of wards (i.e., seats) are randomly reserved for women one election at a time. The study examines two successive local elections held in Mumbai in 1997 and 2002 for the Brihanmumbai (Greater Mumbai) Municipal Corporation (BMC). The author's goal is to use this randomization to test whether the 1997 reservations increased the probability that women contest and win elections in 2002. The author finds that reservations in 1997 approximately doubled the number of female candidates and quintupled the probability of a woman winning office in 2002. We show that these results depend on an assumption that was not mentioned in the original study. There are substantive reasons to believe that the assumption is false, and our reanalysis of the data using an alternative design provides weaker evidence for the effects of quotas.

This natural experiment is unlike the others we discuss in that the assignment of reservations is *known* to be random. In each election, a ward is randomly assigned to be reserved for female candidates or not.[9] The author wants to estimate the effect of wards being reserved for women in 1997 on the probability that women run and are elected in the following

[9] A fixed margins randomization design is used: Both the total number of wards and the number of wards to be reserved are fixed quantities. Wards are randomly selected to be reserved independently and with equal probability. Some scholars, in personal communications with us, claim that political pressure has been used to select which wards are to be reserved, but we, following Bhavnani (2009) and for the sake of our argument, assume randomization.

election, when the quotas are withdrawn. The difficulty with estimating this effect is that a random assignment of reservations occurs again in 2002, which means that the probability that women run and win in 2002 is affected by both interventions. An additional complication is that 30% of wards were reserved for female candidates in 1992, the election prior to 1997.[10]

The author uses a research design that only considers wards that were not reserved in 2002. Wards that are not reserved are also called open wards. In this subsample, the author compares wards that were reserved in 1997 with those that were open in 1997. This design is illustrated in Figure 4(a), where the four possible combinations of reservation assignments in 1997 and 2002 are shown and referred to as A, B, C, and D: Wards that are reserved in 1997 can be either reserved in 2002 (B wards) or open in 2002 (C wards), and similarly, wards that are open in 1997 can be either reserved in 2002 (A wards) or open in 2002 (D wards). Like in our previous two examples, it is possible to define and compare a variety of different treatment and control groups. The most straightforward design compares the 2002 outcomes of wards that in 1997 were reserved (B and C) with wards that were open (A and D). The design proposed by Bhavnani (2009), however, only includes wards that were open in 2002. This design discards A and B wards, and compares C wards (the "treatment" group) to D wards (the "control" group).

The intention behind only keeping wards that are open in 2002 is to mimic a world in which no wards were reserved in 2002. To establish the validity of this design, we must ask whether the difference in the number of women who compete and win in 2002 between C and D wards is the same difference we would have observed *if no wards had been reserved in 2002*. The random assignment of reservations in 1997 and 2002 does not guarantee that these differences are equal. The essential problem is that the design requires that there be no interactions between the two treatment assignments. Because of randomization, *whether* a ward is reserved in 1997 is independent of *whether* it is reserved in 2002. But this independence of treatment assignments does not imply that the effects of the two treatments do not interact with each other. In other words, the design assumes that wards that are open in 2002 would have had the same outcomes if there had been no reservations assigned in 2002. If this assumption is false, and there are substantive reasons to believe that it is, less bias may result if one used all of the wards instead of only the wards that are open in 2002. Our reanalysis of the data shows that whether or not one discards A and B wards is consequential for the substantive findings. If one uses all of the data, the evidence for the effects of quotas is weaker than originally reported.

To help clarify the problem, we develop a hypothetical example where, by assumption, the assignment of reservations in 1997 has no effect on the number of women competing in 2002 in the absence of reservations also being assigned in 2002. Quotas in 1997 do, however, appear to have an effect when reservations are assigned in 2002 because the no–interaction assumption does not hold. Although the example is hypothetical, it is motivated by the secondary literature and accounts in the Indian press that the imposition of quotas may result in fewer women running in open wards. Political parties, which control who may run under their label, may reduce the number of female candidates who run in wards without reservation if there are reservation wards present in the same election. Kishwar (1996, 2872) finds that in local elections with reservations "women are not being allowed to contest from general [open] constituencies which are assumed to be reserved for men." As Kaushik (1992, 49) notes, "Women candidates are viewed as depriving men of their chances." The fact that there are some reserved wards in 2002 influences the number of female candidates who run in open wards in that year.[11]

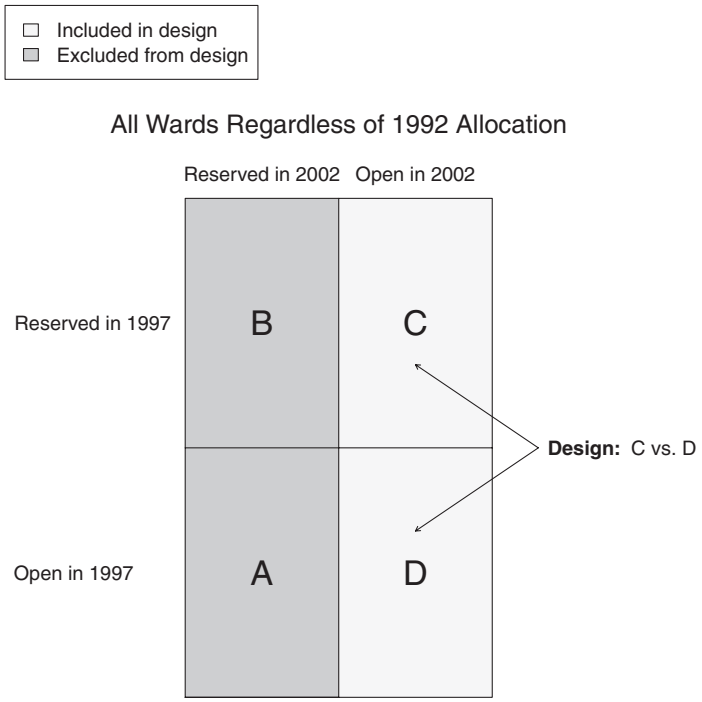
Moreover, if political parties are able to discourage more women from running in 2002 open wards if those wards were also open in 1997, then there is an interaction between the causal effects of the 1997 and 2002 treatment assignments. One reason for why political parties may be better able to discourage women from running in wards that are open in two sequential elections than in wards that were previously reserved is that incumbents and previous candidates may be more difficult to discourage from contesting an election than new candidates. In other words, if a ward is assigned to be open in *both* 1997 and 2002 (cell D in Figure 4(a)), women are more discouraged from contesting the election than if the ward was previously reserved. Thus, in D wards, fewer female candidates are observed in 2002 than would have been observed in the absence of a 2002 reservation allocation. In this example, when we compare C versus D wards, we would observe that the number of women competing in the 2002 election in D wards is lower than the number of women competing in C wards. This would lead us to conclude that 1997 reservations increase womens' electoral outcomes in the following election. However, in this hypothetical example, the difference between treatment and control wards is solely due to a discouragement effect among D wards, wards without reservations in *both* 1997 and 2002. If we repeated the experiment eliminating the allocation of reservations in the second election, we would observe the true zero effect.

We now formally develop this example in the main text and not the Appendix because the formalization is

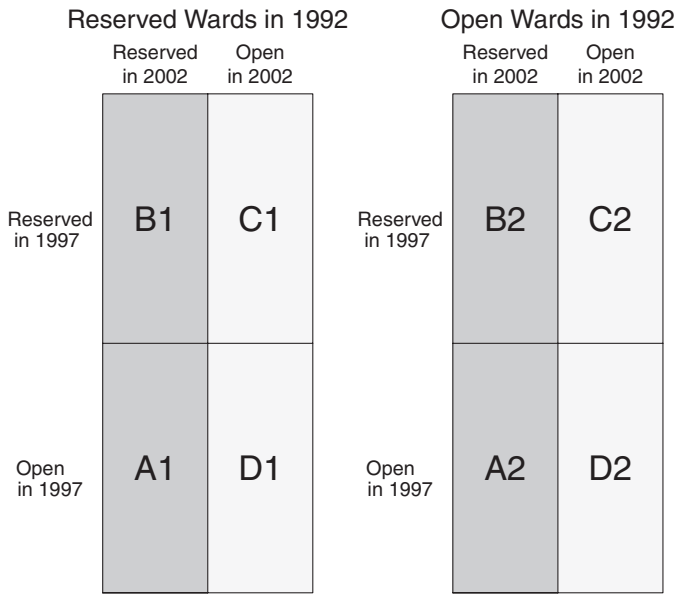
[10] The 1992 reservations were due to a decree by the Government of Maharashtra in March 1990. The later reservations were due to the 73rd amendment to the constitution passed in 1992.

[11] Another reason for fewer women running in open wards with the system of reservations is that women appear to prefer to run against other female candidates. Singh, Lele, Sathe, Sonalkar, and Maydeo (1992), writing about local elections in Maharashtra, note that female candidates are especially vulnerable to slander. Based on their interviews with female candidates, they claim that women feel that female opponents are less likely to engage in character assassination than men. These women prefer to wait for a reserved ward than to challenge men. In the absence of a reservation system, some of these women may be more likely to contest open wards.

**FIGURE 4. Illustration of India Reservations Research Design**




(a) Research design used by Bhavnani (2009), ignoring 1992 reservation allocations.




(b) Research design used by Bhavnani (2009), considering 1992 reservation allocations.

straightforward. The units are wards and there are two treatments: $T_{97,i} = 1$ if ward $i$ is reserved for women in 1997, and $T_{97,i} = 0$ if the ward is open in 1997. Similarly, $T_{02,i} = 1$ if ward $i$ is reserved for women in 2002, and $T_{02,i} = 0$ if the ward is open in 2002. $Y_{02,i}$ is the observed outcome in 2002. In 2002, there are four possible (i.e., potential) outcomes, one for each possible combination of treatment assignments in 1997 and 2002, which we refer to by $Y_{02}(T_{97}, T_{02})$, after dropping $i$ subscripts. The four possible outcomes correspond to the four combinations of treatments that can occur, as shown in Figure 4(a).

$T_{97}$ and $T_{02}$ are randomly and independently assigned in 1997 and 2002, respectively. The observed outcome in 2002 is

$$\begin{aligned} Y_{02} = {} & T_{97} \cdot T_{02} \cdot Y_{02}(1,1) + (1 - T_{97}) \cdot T_{02} \cdot Y_{02}(0,1) \\ & + T_{97} \cdot (1 - T_{02}) \cdot Y_{02}(1,0) \\ & + (1 - T_{97}) \cdot (1 - T_{02}) \cdot Y_{02}(0,0). \end{aligned} \quad \textbf{(1)}$$

In a world where no randomization took place in 2002, the potential outcomes in 2002 would only depend on the treatment assigned in 1997. We denote these potential outcomes by $\widetilde{Y}_{02}(1)$ and $\widetilde{Y}_{02}(0)$, respectively, for $T_{97} = 1$ and $T_{97} = 0$. The effect of interest is $\widetilde{\tau} = E(\widetilde{Y}_{02}(1) - \widetilde{Y}_{02}(0))$, but equation 1 shows that this effect cannot be recovered by this natural experiment unless we assume that the 2002 randomization does not change any of the potential outcomes—i.e., $Y_{02}(1,1) = Y_{02}(1,0) = \widetilde{Y}_{02}(1)$ and $Y_{02}(0,1) = Y_{02}(0,0) = \widetilde{Y}_{02}(0)$.[12] In other words, we can recover the treatment effect of interest if there are no interactions between the 1997 and 2002 treatments.

We now turn to the hypothetical example discussed earlier. We assume that there is no treatment effect of 1997 quotas, $\widetilde{\tau} = 0$, but that the potential outcomes in the world where both randomizations occur are

$$Y_{02}(1,1) = Y_{02}(0,1) \quad \textbf{(2)}$$

$$Y_{02}(1,0) = \widetilde{Y}_{02}(1) \quad \textbf{(3)}$$

$$Y_{02}(0,0) = \widetilde{Y}_{02}(0) - K \quad \textbf{(4)}$$

where $K$ is a constant such that $K > 0$.

Equation 2 says that when a ward receives the treatment in 2002, its outcome is the same regardless of what treatment was received in 1997. This might be the most plausible assumption because a ward that is reserved in 2002 must elect a women regardless of what happened in 1997. But even in this case, a substantive assumption is being made. For example, perhaps there would be more female candidates in wards that were reserved two elections in a row than simply in the current election.

Equation 3 says that, when the treatment is received in 1997 and not in 2002, the outcome is the same as the outcome that would have been observed in a world where the 2002 randomization had not existed and the ward had been assigned to treatment in 1997. Equation 4 says that there is a discouragement effect: When the ward is assigned to control in both 1997 and 2002, fewer women run for office than the number of women who would have run in a world where only the 1997 randomization had existed and the ward had been assigned to the open condition in this randomization. Note that equation 3 assumes that there is no discouragement effect in 2002 if a ward was reserved in 1997. This is a simplification to make the math clear. Likewise, equation 2 is stricter than necessary. As noted earlier, what is needed for this example is that the parties have more power to discourage women from running when wards are assigned to be open in both 1997 and 2002 than when they are only assigned to be open in 2002.

The author's design keeps wards that are open in 2002 ($T_{02} = 0$), and then compares wards that were reserved in 1997 ($T_{97} = 1$) to wards that were open in 1997 ($T_{97} = 0$). The parameter estimated by this design is

$$\begin{aligned} & E(Y_{02} \mid T_{97} = 1\,,\ T_{02} = 0) \\ & \quad - E(Y_{02} \mid T_{97} = 0\,,\ T_{02} = 0) \\ & = E(Y_{02}(1,0) \mid T_{97} = 1\,,\ T_{02} = 0) \\ & \quad - E(Y_{02}(0,0) \mid T_{97} = 0\,,\ T_{02} = 0) \\ & = E(Y_{02}(1,0)) - E(Y_{02}(0,0)) \\ & = \widetilde{\tau} + K > \widetilde{\tau}. \end{aligned}$$

In this counterexample, the design overestimates the true treatment effect $\widetilde{\tau}$. Even when $\widetilde{\tau} = 0$, the design would estimate a spurious positive effect equal to $K$, which is the discouragement effect. In other words, comparing C and D wards leads to a misleading answer. Could the problem be avoided by including A and B wards in the analysis? If no wards are discarded, the analysis now compares B and C wards combined (both reserved in 1997) to A and D wards combined (both open in 1997). As we show in the Appendix, including all wards in the analysis may lead to an estimate that is smaller and closer to the true effect than the effect estimated when A and B wards are discarded (although this effect would still be partially contaminated by the discouragement effect). We prove that in this example the bias using the entire population will be less than in the subpopulation with $T_{02} = 0$ for all $\widetilde{\tau}$ such that $-K < \widetilde{\tau} < K(2 - p_{02})/p_{02}$, where $p_{02}$ is the probability of receiving treatment in 2002. This probability is $\frac{1}{3}$ in our data. The intuition is that using all of the data includes wards (A and B) that are not contaminated by an interaction between the two treatments. Therefore, the average effect over all wards will be less biased, because it is a combination of a biased comparison (C and D) and an unbiased comparison (A and B).

It is possible to construct many other examples in which discarding wards increases bias. But it is also possible to construct examples where discarding A and B wards is beneficial, in the sense of yielding an effect closer to the true zero effect. The crucial point is that it is impossible to decide whether or not to subset the data without specifically assuming how the effects of the 1997 and 2002 allocations of reservations interact with one another. In other words, the fact that a treatment (reservations) is assigned in 2002, as well as in 1997, introduces the possibility of interactions between both effects. Therefore, to use this design to estimate the effect of 1997 reservations *alone* on 2002 outcomes, one must *assume* whether and how the effect

[12] Note that, in both cases, only the first equality is an assumption, and the second is just a redefinition. Also, this is a sufficient condition but it is not necessary (equality of means would suffice for $\widetilde{\tau}$).

**TABLE 5. Next-election Effects of the 1997 Reservations on the 2002 Elections**

| | Using Only Wards Open in 2002 | | | | Using All Wards | | | |
|---|---|---|---|---|---|---|---|---|
| | Reserved '97 | Open '97 | Diff. | *p*-Value | Reserved '97 | Open '97 | Diff. | *p*-Value |
| Percentage of female winners | 21.62 | 3.70 | 17.92 | 0.00 | 47.27 | 33.90 | 13.37 | 0.10 |
| Percentage of wards where at least one woman ran for office | 72.97 | 35.80 | 37.17 | 0.00 | 81.82 | 55.93 | 25.89 | 0.00 |
| Number of female candidates | 1.14 | 0.46 | 0.68 | 0.00 | 3.18 | 2.36 | 0.82 | 0.15 |
| Number of candidates | 10.59 | 9.14 | 1.46 | 0.14 | 9.55 | 8.32 | 1.22 | 0.10 |
| Female candidates as a percentage of candidates | 11.86 | 4.41 | 7.45 | 0.00 | 40.70 | 34.38 | 6.32 | 0.40 |
| Number of competitive female candidates | 0.46 | 0.14 | 0.32 | 0.00 | 1.53 | 1.23 | 0.30 | 0.33 |
| Number of competitive candidates | 4.14 | 3.91 | 0.22 | 0.31 | 4.14 | 3.91 | 0.22 | 0.33 |
| Female percentage of competitive candidates | 11.83 | 3.19 | 8.64 | 0.00 | 40.68 | 33.55 | 7.14 | 0.35 |
| Number of new female candidates | 0.07 | 0.03 | 0.04 | 0.01 | 0.31 | 0.31 | 0.01 | 0.91 |
| Percentage of wards where any female candidate was competitive | 43.24 | 13.58 | 29.66 | 0.00 | 61.82 | 40.68 | 21.14 | 0.01 |
| Total percentage of votes for female candidates | 15.04 | 3.35 | 11.69 | 0.00 | 42.84 | 33.65 | 9.19 | 0.22 |
| Average percentage for female candidates | 9.96 | 2.43 | 7.53 | 0.00 | 11.79 | 7.47 | 4.33 | 0.03 |
| Turnout | 41.57 | 42.20 | −0.62 | 0.61 | 41.65 | 41.76 | −0.11 | 0.88 |
| Winning candidate vote percentage | 40.96 | 42.82 | −1.86 | 0.38 | 43.00 | 43.49 | −0.49 | 0.81 |
| Winning candidate vote margin | 13.49 | 15.29 | −1.80 | 0.50 | 15.71 | 15.25 | 0.46 | 0.80 |
| Number of wards | 37 | 81 | | | 55 | 118 | | |

*Notes*: All hypothesis test are permutation tests consistent with the known randomization (i.e., both the total number of wards and the number of wards to be reserved are fixed quantities). Wards are randomly selected to be reserved independently and with equal probability.

of 1997 reservations on 2002 outcomes changes when 2002 reservations are introduced.

In our example, the estimated treatment effect using all of the data is smaller than the estimated effect using only wards that are open in 2002. We find this to be the case in the actual data. Table 5 presents the results for both the Bhavnani design and the one that uses all of the wards in 2002.[13] If one discards A and B wards and restricts the analysis to only wards that were open in 2002, then the effects of 1997 quotas in 2002 appear to be strong. Wards that were reserved in 1997 are, in 2002, significantly more likely to have a women win (22% vs. 4%), have at least one female candidate (73% versus 36%), have more female candidates (1.1 vs. .46), have a greater percentage of candidates be women (12% versus 4%), have more new female candidates who did not run in 1997 (0.07 versus 0.03), have competitive[14] female candidates (43% vs. 14%), and have higher average vote percentages for female candidates (10% vs. 2%). In contrast, if all wards are used, the only results that are clearly significant are the percentage of wards where at least one woman ran for office (82% vs. 56%), the percentage of wards where any female candidate was competitive (62% vs. 41%), and the average percentage for female candidates (12% vs. 7%).

We now turn to the issue of the 1992 election. We show that if the 1992 reservations had the same effect as originally reported for the 1997 reservations, then the number of winning female candidates observed in 1997 is anomalously low. This finding then provides further evidence for a smaller treatment effect than originally reported and implicitly against the no–interaction assumption made by the original paper.[15]

The exact method by which wards were reserved in 1992 is unknown, but it may not have been random. But even if the reservations in 1992 were not randomly assigned, whether or not a ward was reserved in 1992 is independent of the assignments in 1997 and 2002 because the later two were randomized. The author acknowledges that the 1992 reservations did occur, but does not use them because of the uncertainty over whether they were randomly assigned and because of significant changes in the ward boundaries between 1992 and 1997. The author clearly states that the estimates he reports are conditional on there having been reservations in 1992, but otherwise ignores the issue. However, if there are reasons to believe that the effect of 1997 reservations on 2002 should be restricted to open wards in 2002, the same reasons would hold for 1992 reservations because their assignment is independent of the later reservation assignments.

The 1992, 1997, and 2002 interventions are illustrated in Figure 4(b), where wards are classified according

[13] Our Table 5 replicates and extends Bhavnani's table 3 (2009).

[14] Following Bhavnani (2009), a candidate is considered competitive if she receives at least 5% of the vote.

[15] Of course, it may also be possible that the estimates from 1997 simply do not hold in 1992 because of problems of external validity.

to their reservation status. Once we incorporate 1992 reservations, we can divide the author's original comparison (C vs. D in Figure 4(a)) into two subcomparisons: (1) for the subset of wards that are reserved in 1992, we compare wards that are reserved in 1997 and open in 2002 (cell C1 in Figure 4(b)) to wards that are open in 1997 and open in 2002 (cell D1 in Figure 4(b)); and (2) for the subset of wards that are open in 1992, we compare wards that are reserved in 1997 and open in 2002 (cell C2 in Figure 4(b)) to wards that are open in 1997 and open in 2002 (cell D2 in Figure 4(b)).

If the next-election effects of reservations are as large in 1992 as they were originally reported to be in 1997, we should see that the number of women who dispute and win office in 1997 is larger in D1 wards (which are reserved in 1992 but open in 1997) than in D2 wards (which are open in both 1992 and 1997). The design proposed by the author pools D1 and D2 wards (leading to D wards in Figure 4(a)). If 1992 reservations have an effect on 1997 outcomes, pooling wards in this way will contaminate the starting level of female candidacies in 1997. In light of this, if the next-election effect of winning is significant and 18% as originally reported, it is surprising that the percentage of female winners in open wards in 1997 is only 3.4% (table 2 in Bhavnani 2009). If the original estimates in Table 5 hold, the number of female winners in open 1997 wards should be about 9%, and if this is the correct estimate, observing a value equal to or lower than 3.4% is highly unlikely ($p$-value $= 0.01$).[16] Recall that when we use all of the data, we do not find a significant effect for women winning reserved wards in the next election.

In conclusion, it is not clear whether one should analyze all of the data or only the subsample in which there were no reservations in 2002. We think there are substantive reasons to prefer to analyze all of the data, but others may disagree. For our purposes, the key point is that even in this example with actual randomization, the two questions we have highlighted in this article are important to ask: Is the proposed comparison valid under random assignment? If not, what is the comparison that is guaranteed by the randomization, and how does this comparison relate to the comparison the researcher wishes to make? For the estimand of interest, one is required to make an assumption not guaranteed by the randomization itself: the no–interaction assumption. Given this complication, we suggest treating the 2002 reservations like the author treats the 1992 reservations: Use all of the data and explicitly state that the experimental results are conditional on having reservations in 1992 and 2002.

[16] Hypothesis test conducted using permutation inference. Approximately 30% of wards that are open in 1997 were reserved in 1992. Therefore, $0.09 \approx 0.3 * 0.216 + (1 - 0.3) * 0.037$, where 0.216 and 0.037 are, respectively, the originally reported estimates of the proportion of wards with female winners in wards that were previously treated and wards that were previously open.

## A NATURAL EXPERIMENT BASED ON A DISCONTINUITY

We now discuss two issues concerning the application of regression discontinuity (RD) to estimate the incumbency advantage. Lee (2008) proposed a design to estimate the incumbency advantage in the U.S. House based on the discontinuity that occurs at the 50% vote share cutoff in a two-party system: The party that obtains a vote share equal to or above 50% wins the election, whereas the party that obtains a vote share below 50%, no matter how close to 50%, loses the election. Thus, one can think of winning the election as being as-if randomly assigned among districts where the party obtains a vote share very close to 50%. The effect of interest in this design is the impact of a party winning the election at time $t$ on the vote share obtained by the party in the following election, at $t + 1$. The treatment of interest is winning at $t$; that is, becoming the incumbent party at $t$. The outcome of interest is the vote share obtained at $t + 1$. The treatment group is composed of districts where the party barely won at $t$ (also called *bare-winner districts*), and the control group is composed of districts where the party barely lost at $t$ (also called *bare-loser* districts).[17] There is evidence that parties and candidates may have the ability to sort around the 50% vote share cutoff in U.S. House elections (Caughey and Sekhon 2011). We ignore this issue and assume that winning or losing is randomly assigned in close elections so that we can analyze the design in its best-case scenario. If this assumption is satisfied, a comparison between the vote shares obtained by the Democratic (or Republican) Party at $t + 1$ in bare-winner and bare-loser districts will give us a valid estimate of the impact of the party's winning at $t$ on the party's vote share at $t + 1$.

The first issue we analyze is the study by Butler (2009) that examines the impact of tenure on the incumbency advantage using RD to estimate the difference between freshmen and nonfreshmen incumbents in their general election vote shares. In this study, the effect of interest is the change in the incumbency advantage that is caused by being a freshman as opposed to being an incumbent who has already served at least two periods. Because the treatment of interest is whether the incumbent is a freshman or not, the author makes a fundamental modification to the original RD design described earlier: All districts where an incumbent is not running in election $t$ or $t + 1$ are dropped. These restrictions create the following scenario. Because there are no open seats at $t$, there is always an incumbent running for reelection at $t$. In those districts where the incumbent at $t$ barely loses, some challenger barely wins, and this challenger will be a freshman incumbent when running for reelection at $t + 1$. In those districts where the incumbent at $t$ barely wins, the winner will not be a freshman at $t + 1$, because he or she will have been an incumbent for at

[17] This is an intuitive albeit loose definition of treatment and control groups in an RD design. For technical details, see Hahn, Todd, and van der Klaauw (2001).

least two terms and possibly longer. Moreover, everyone elected at $t$ is observed again at $t+1$, because all districts where incumbents do not run again in election $t+1$ are dropped from the sample.

Imposing these restrictions generates a redefinition of the treatment in an important way. Originally, the randomly assigned treatment is winning. The author, however, is not interested in the effect of winning, but rather in the effect of an attribute of winners and losers—being a freshman. Yet, even if close elections are decided randomly, whether the winner is a freshman is not randomly assigned. To address this issue, the author imposes the restrictions outlined earlier, and as a result the attribute of interest (being a freshman or not) and the treatment (who wins) overlap perfectly.

The crucial question is whether these modifications are valid under random assignment or require additional assumptions. Does randomization justify dropping open seats at $t$ and $t+1$ from the analysis? It does not. Assuming winners are randomly assigned guarantees that all bare-winner districts are comparable to all bare-loser districts, but it does not guarantee that the observed difference between these groups in the subset of the data that discards open seats is the true effect of tenure. This is partly analogous to our previous example on mandated gender quotas, in which the random assignment of reservations alone was not enough to guarantee that keeping a subset of the data would lead to the true effect of reservations.

In this case, there are two issues to be distinguished: dropping open seats at $t$, and dropping open seats at $t+1$. Assuming that open seats at $t$ are decided before the randomization takes place, keeping only incumbent-held districts at $t$ will not affect the comparability of the treatment and control groups.[18] Restricting to incumbent-held seats at $t+1$, however, is problematic for at least two reasons. First, incumbents who barely defeated a challenger at $t$ (the treatment group) might be seen as more vulnerable at $t+1$ than challengers who barely defeated an incumbent at $t$ (the control group) and as a result might be more likely to be targeted by the opposite party at $t+1$. This will affect the vote shares of the treatment group but not those in the control group, so that the comparison of both groups at $t+1$ may lead to biased results. Second, incumbents might decide to retire before $t+1$ in anticipation of a bad electoral outcome. And if part of this anticipated negative outcome carries over to the party's new candidate, dropping open seats will lead to a sample in which vote shares are higher than would have been observed in the entire sample. If incumbents are more likely to retire strategically the longer they have held the seat, the treatment group (which is composed of only nonfreshman incumbents and discards all seats where these incumbents decide to retire) will tend to have higher vote shares than the control group, and a comparison of both groups will again lead to incorrect results.

In fact, Caughey and Sekhon (2011) show that in close House elections from 1948–2008, 20% of incumbents who barely win reelection retire in the next election, whereas no victorious challengers who barely win retire. These differing retirement rates are consistent with strategic behavior. Restricting attention to seats that are closed at $t+1$ when one-fifth of incumbents retire but no challengers do makes it unlikely that the remaining incumbents and challengers are comparable. As explained earlier, it is likely that the stronger incumbents remain, which would result in the author's estimates being larger than they should be.

Once again, we have shown that randomization alone does not guarantee that the design recovers the true effect. In this case, additional assumptions about the strategic behavior of legislators are needed. These assumptions are not explicitly made by the author, and the available evidence on strategic retirement suggests that they do not hold in the case of the U.S. House.

Next, we discuss the validity of RD applications that pool observations across two or more elections. This has been the norm in all empirical applications of this design to measure the incumbency advantage in the U.S. House (including the article by Butler [2009] just analyzed), which usually pool all general elections after 1948. The random assignment of winners and losers in close elections, however, does not guarantee that this pooling across time leads to the effect of interest for most researchers. Because it focuses on close races, in any given election the RD design estimates a *local* treatment effect; that is, the effect of winning for the subsets of districts with close races in that election. For example, the RD estimate of the effect of the Democratic Party winning in 1994 on the 1996 Democratic vote shares is the effect of winning for the type of districts that have close races in 1994. Similarly, the RD estimate of the effect of the Democratic Party winning in 2006 on the 2008 Democratic vote shares is the effect for the type of districts that have close races in 2006. But these two effects might be different, because districts with close races in 1994—a bad year for the Democratic Party—might differ from districts with close races in 2006—a bad year for the Republican Party. The cyclical nature of elections–in particular the phenomenon of midterm loss–might make these local effects very different from one another.

It follows that the average effect in the pooled data will potentially mask great heterogeneity, and it may fail to be representative of any particular election. Nonetheless, scholars have often been interested in obtaining the average effect of incumbency over long periods of time, and RD analyses that pool across many elections are motivated by an interest in that average effect. Unfortunately, even if winning is randomly assigned in close elections, using the RD design in an analysis that pools observations across elections will not necessarily recover the average of all the single-election effects. The reason is that the number of bare-winner and bare-loser districts changes over time. Calculating the overall average in the pooled data

[18] We may observe that the effect among incumbent-held seats at $t$ is very different from the effect among open seats at $t$ or among all seats. Nonetheless, in the subsample of incumbent-held seats at $t$, treatment and control groups will be comparable.

implicitly weighs the effect of incumbency differently from the way in which this effect is weighted if one first calculates individual average effects for every election and then takes the average of all these local (i.e., election-specific) averages. As we show formally for the case of two successive elections in the Appendix, a sufficient condition for pooling data across elections to recover this average of the local effects is that the number of bare-winner districts is constant in every election, and similarly for bare-loser districts. This condition is not guaranteed to hold by the random assignment of winners in close elections, because both the probability of winning and the total number of districts with close races may change from one election to the next even under random assignment. In fact, these numbers do change considerably in U.S. House elections over time, which implies that a pooled analysis of congressional elections does not yield the average of all single-election incumbency effects.

This case is unlike our previous examples in that, considering one election at a time, there are only two possible groups to be compared (bare winners and bare losers), and it is straightforward to define one as treatment and the other as control. Moreover, the assumed randomization guarantees that they are comparable (unless, as Butler's example shows, one subsets the data in ways that compromise this comparability). Subtleties and possible complications arise when one considers multiple elections and mistakenly treats the data as having come from a single large experiment instead of considering it as a collection of multiple individual experiments. Pooling the data collapses the election-specific treatment groups into a single treatment group, and the election-specific control groups into a single control group. But this is an extra step in the analysis, unrelated to the assumptions behind the RD design itself. As we show in the Appendix, there are conditions under which a pooled analysis leads to the same effect as the average of all individual effects, but these conditions are not guaranteed to hold under random assignment.

## CONCLUSION

It is often said that with a good research design, one does not need statistics. As we show in this article, one needs statistics to evaluate the research design, to know if it is indeed good. Although natural experiments offer significant advantages, they do not possess key benefits of actual experiments and hence require careful theoretical and statistical work to make valid inferences. First, there is the obvious problem that natural experiments seldom have a known random assignment mechanism. As a consequence, and as seen in the redistricting example, rarely can natural experiments be used without significant covariate adjustment.

A less often noted but crucial issue with natural experiments is that the treatment and control groups researchers want to use may not be comparable even if one assumes random assignment. Why do we not see this phenomenon often in controlled randomized experiments? The answer lies in the characteristic feature of natural experiments: Natural experiments are situations in which an intervention is—in the best-case scenario—randomly assigned, but this intervention is not under the control of the researcher. Rather, it is assigned by nature or by individuals whose goals differ from those of the researcher. In contrast, when designing a controlled randomized experiment, researchers a priori design the study so that randomization will ensure that treatment and control groups of interest are comparable. The result is that, although distinguishing treatment from control groups is straightforward in randomized experiments, establishing which groups ought to be compared is much more complicated in natural experiments.

Our analysis of four natural experiments used in different areas of political science shows that a successful strategy is to first ask what comparable treatment and control groups would have been created if the natural intervention had been randomly assigned, and then to ask if and how comparing these groups is related to the causal effect of interest. This strategy makes clear that a natural experiment whose intervention is random can still lead to incorrect inferences for two reasons: The groups that the researcher wishes to compare are not the groups that the intervention guarantees to be comparable, or the groups that the intervention guarantees to be comparable are not the groups that are informative about the effect of interest. Because the design of the natural manipulation is not controlled by the researcher, the valid comparison is often far removed from the comparison that would be most informative for the substantive question the researcher is studying. This creates a tension between internal validity and substantive relevance.

The example of redistricting to estimate the personal vote illustrates this tradeoff well. Unlike a comparison of New Neighbors and New Voters, a comparison of Late and Early New Voters is valid under the assumption of random redistricting, but when applied to successive redistricting plans, this comparison only captures the portion of the personal vote that accrues in two years. If some amount of goodwill toward the incumbent takes longer to accumulate, it will not be part of the effect estimated by this design. This is the price we must pay if we want to use redistricting to estimate the personal vote under the weakest assumptions. We were faced with a similar tradeoff when we considered the available designs under a single redistricting plan. The one-time redistricting design that compares Old Neighbors to New Voters is guaranteed to be valid under randomization. However, these groups are in a different district after redistricting, thus their comparison is very weakly related to the personal vote. In this case, too much relevance had to be given up to gain validity, and the design had to be abandoned.

The issues we discuss have received scant attention in the methodological literature. A related but distinct idea is design sensitivity (Rosenbaum 2004, 2010), which is a measure that compares the relative effectiveness of competing designs in large samples. For a

specific treatment effect and research design, design sensitivity asks how much unobserved bias would be needed to make the null hypothesis of no treatment effect plausible. This type of analysis is used to judge the ability of different features of the research design, such as the number of control groups, the number of outcomes that are expected to be affected by the treatment, and the dose in which the treatment is given, to discern treatment effects from bias due to unobserved covariates.

The framework we propose is best conceived as a complementary strategy to design sensitivity analysis, but one that comes before it. Design sensitivity analysis varies certain features of the research design, but leaves fixed the definitions of the treatment and control conditions, because it simply assumes that these definitions are uncontroversial. Our framework invites researchers to reflect on these definitions in the first place, to establish the assumptions that would be required to characterize a group as treatment and another as control, and to determine whether these assumptions are guaranteed to hold in the best-case scenario of random assignment.

Finally, we do not intend our argument to imply that natural experiments have no advantages over other research designs. Our framework is only possible because natural experiments have a clear intervention (e.g., a redistricting plan, an election winner or loser, a move from the House to the Senate). This is one of the benefits of natural experiments: Even when the as-if random assumption is false, one can think through the process by which the intervention was assigned and how the precise intervention relates to the substantive question at hand. More generally, with such studies, it is easier to determine what is post- and what is pre-treatment than with the observational designs more commonly used and certainly easier than with cross-sectional data that lack any intervention. In addition, placebo tests may be available to help determine if controlling for observable characteristics increases the plausibility of the as-if randomness assumption.

## APPENDIX: FORMALIZATION OF EXAMPLES

We develop formally the examples we discuss in the main text using the potential outcomes framework (Holland 1986; Rubin 1974).[19] We do not formally discuss our second example on the voting behavior of U.S. House members who move to the Senate, because it is analogous to the redistricting example and its formalization is similar.

In the potential outcomes framework, units are assigned a binary treatment, and every unit is thought of as having two *potential* outcomes: one outcome that occurs if the unit receives the treatment condition and another outcome that occurs if the unit receives the control condition. Our notation is as follows. Every unit $i$ receives a treatment $T_i$, with $T_i = 1$ if $i$ receives the treatment condition and $T_i = 0$ if $i$ receives the control condition. Unit $i$'s potential outcome under treatment is $Y_{1i}$, and its potential outcome under control is $Y_{0i}$. The observed outcome is denoted $Y_i$ and is defined as $Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i}$. The so-called fundamental problem of causal inference is that for every unit $i$ we either observe $Y_{1i}$ or $Y_{0i}$, but never both. The individual-level effect is $\tau_i = Y_{1i} - Y_{0i}$, but because of this fundamental problem this effect is not identified. In observational studies, inference is usually about the expectation of $\tau_i$ taken over some subset of the units. We use this general notation, with appropriate modifications, to formalize certain aspects of the examples discussed in the main text.

## REDISTRICTING AND THE PERSONAL VOTE

To formalize the discussion in the section *Redistricting as a Natural Experiment for the Personal Vote*, let $T_i$ be equal to 1 if precinct $i$ is moved from one district to another just before election $t$ and equal to 0 if precinct $i$ is not moved to a different district before election $t$, and let $D_i$ be equal to 1 if precinct $i$ has new voters in its district at election $t$ and equal to 0 if precinct $i$ has no new voters in its district at election $t$. Let $Y_0(i, t)$ be the incumbent vote share in precinct $i$ if $T_i = 0$ and $D_i = 0$ (the precinct is not moved and does not have new neighbors, i.e., these are voters who stay in $A$ after redistricting), let $Y_1(i, t)$ be the incumbent vote share in precinct $i$ if $T_i = 0$ and $D_i = 1$ (the precinct is not moved and has new neighbors, i.e., these are voters who are in $B$ before and after redistricting), and let $Y_2(i, t)$ be the incumbent vote share in precinct $i$ if $T_i = 1$ and $D_i = 1$ (the precinct is moved and has new neighbors, i.e., these are voters who are moved from $A$ to $B$).[20] Of course, the fundamental problem of causal inference is that for every precinct we observe only one of its three potential outcomes. This is, we only observe the *realized* incumbent vote share, defined as

$$Y(i, t) = Y_0(i, t) \cdot (1 - T_i) \cdot (1 - D_i) + Y_1(i, t) \cdot (1 - T_i) \cdot D_i + Y_2(i, t) \cdot T_i \cdot D_i. \quad \textbf{(5)}$$

As is common with observational studies, we focus on the average treatment effect on the treated (ATT). Given the setup of our hypothetical experiment, the ATT can be defined in two different ways:

$$ATT_0 \equiv E[Y_2(i, t) - Y_0(i, t) \mid T_i = 1, D_i = 1]. \quad \textbf{(6)}$$

$$ATT_1 \equiv E[Y_2(i, t) - Y_1(i, t) \mid T_i = 1, D_i = 1]. \quad \textbf{(7)}$$

It can be shown that the following condition is sufficient for $ATT_0$ to be identified[21]:

$$E[Y_0(i, t) \mid T_i = 1, D_i = 1] = E[Y_0(i, t) \mid T_i = 0, D_i = 0]. \quad \textbf{(8)}$$

Similarly, it can be shown that the following condition identifies $ATT_1$:

$$E[Y_1(i, t) \mid T_i = 1, D_i = 1] = E[Y_1(i, t) \mid T_i = 0, D_i = 1]. \quad \textbf{(9)}$$

[19] See Sekhon (2008) for a review of this framework and its origins.

[20] The potential outcome when $T_i = 1$ and $D_i = 0$ is not defined because it is not possible to be moved from one district to another and not to have new neighbors.

[21] For a formal treatment of these and related assumptions, see, for example, Heckman, Ichimura, and Todd (1998).

In words, Assumption (8) says that voters who stay in $A$ and voters who are moved from $A$ to $B$ would have attained the same average outcomes if they had not been moved and if they had not received new neighbors in their districts. Assumption (9), in contrast, states that voters who are originally in $B$ and voters who are moved from $A$ to $B$ would have attained the same average outcomes if $A$'s voters had not been moved and $B$'s voters had not received new neighbors.

This makes clear that randomization does not imply that $B$'s old voters are a valid counterfactual for $B$'s new voters: although randomization, if successful, ensures that Assumption (8) is satisfied (and hence that the average treatment effect defined by Equation (6) is identified), randomization does *not* imply Assumption (9). In other words, randomization ensures exchangeability between the set of voters for which $(1 - T_i) \cdot (1 - D_i) = 1$ (i.e., voters who stay in $A$ after redistricting) and the set of voters for which $T_i \cdot D_i = 1$ (i.e., voters who are redistricted from $A$ to $B$), but not between the latter set of voters and the set of voters for which $(1 - T_i) \cdot D_i = 1$ (i.e., voters who are originally in $B$).

We now consider additional methodological issues that arise if one decides to implement the second-best one-time redistricting design despite its difficulties. Because Assumption (9) is not valid even with random assignment, we define a weaker version of this assumption:

$$\begin{aligned} &E\left[Y_1\left(i,t\right) \mid T_i = 1, D_i = 1, X\right] \\ &\quad = E\left[Y_1\left(i,t\right) \mid T_i = 0, D_i = 1, X\right], \end{aligned} \tag{10}$$

where $X$ is a vector of observable characteristics. Assumption (10) can be shown to identify $ATT_1$ conditional on $X$ and is considerably weaker than Assumption (9). Thus, if one is still interested in using $B$'s original voters as counterfactuals despite the methodological difficulties, one could attempt to find the subpopulation of $B$'s old voters who are most similar to the new voters on some set $X$ of observable characteristics and use these as counterfactuals, under the assumption that once the joint distribution of $X$ is equated among new voters and new neighbors, their average potential outcomes would have been identical in the absence of redistricting. But note that Assumption 10 defines a selection on observables assumption that is not guaranteed to hold even under random assignment!

To complicate things further, if Assumption 10 were true this approach may still not result in unbiased estimates, because the distribution of $X$ between $B$'s old and new voters is not guaranteed to be equal *even if conditional on X both groups of voters would have attained the same average outcomes in the absence of redistricting*.[22] The reason is that the support of the distribution of $X$ among $B$'s new voters may be different from the support of the distribution of $X$ among $B$'s old voters, a concern that becomes all the more relevant given that $B$'s old and new voters were originally in different districts.

## Consecutive Redistricting

To formally establish the parameter identified by the FBTT design, let $W_{i,t+1} = 1$ if precinct $i$ is moved from district $A$ to district $B$ at election $t+1$, and $W_{i,t+1} = 0$ if precinct $i$ is moved from $A$ to $B$ at election $t$ and remains in $B$ at election $t+1$. In other words, $W_{i,t+1}$ is a late new voter treatment indicator, where late new voter is defined as voting in $B$ for the first time at election $t+1$. Letting $Y_0\left(i, t+1\right)$ denote the incumbent vote share in $i$ at election $t+1$ if $W_{i,t+1} = 0$ and letting $Y_1\left(i, t+1\right)$ denote the incumbent vote share in $i$ at election $t+1$ if $W_{i,t+1} = 1$, we define the parameter of interest, $ATT_2$, as

$$ATT_2 \equiv E\left[Y_1\left(i, t+1\right) - Y_0\left(i, t+1\right) \mid W_{i,t+1} = 1\right], \tag{11}$$

which is identified under

$$E\left[Y_0\left(i, t+1\right) \mid W_{i,t+1} = 1\right] = E\left[Y_0\left(i, t+1\right) \mid W_{i,t+1} = 0\right]. \tag{12}$$

In words, $ATT_2$ is identified if late new voters and early new voters would have attained the same average outcomes if they both had been in the new district for exactly two elections. Later, we show that randomization plus a stationarity assumption guarantee that Assumption (12) holds.

Because we assumed that both groups of voters are in the same district at election $t-1$, and that just before election $t$ the set of voters for which $W_{i,t+1} = 0$ is randomly chosen and moved to district $B$, we have

$$E\left[Y_0\left(i, t-1\right) \mid W_{i,t+1} = 1\right] = E\left[Y_0\left(i, t-1\right) \mid W_{i,t+1} = 0\right]. \tag{13}$$

Assumption (13), implied by randomization, guarantees that both groups of voters have the same pre-treatment average outcomes at $t-1$. But Assumption (13) does not imply Assumption (12); hence we need to add an assumption to the FBTT design to obtain exchangeability at election $t+1$. We make the following additional assumption:

$$\begin{aligned} &E\left[Y_0\left(i, t+1\right) - Y_0\left(i, t-1\right) \mid W_{i,t+1} = 1\right] \\ &\quad = E\left[Y_0\left(i, t+1\right) - Y_0\left(i, t-1\right) \mid W_{i,t+1} = 0\right]. \end{aligned} \tag{14}$$

Assumptions (13) and (14) together imply Assumption (12). If late new voters are randomly chosen *and* early new voters and late new voters would have followed the same path between election $t-1$ and election $t+1$ if they both had spent election $t$ and election $t+1$ in the new district, $ATT_2$ is identified. As before, because district boundaries are not randomly modified, in practice we modify Assumptions (13) and (14) to make them conditional on $X$.

## MANDATED QUOTAS AND WOMEN'S PROBABILITY OF WINNING

Following our discussion in the section *Randomized Gender Quotas in India*, we show that there are a range of values of $K$ and $\tilde{\tau}$ for which using the entire population leads to a *smaller* bias than using only the subpopulation with $T_{02} = 0$. To see this, let $E\{T_{02}\} = p_{02}$ be the probability of receiving treatment in 2002, and observe that

$$\begin{aligned} E\{Y_{02} \mid T_{97} = 1\} &= E\{T_{02} \cdot Y_{02}(1,1) \mid T_{97} = 1\} \\ &\quad + E\{(1 - T_{02}) \cdot Y_{02}(1,0) \mid T_{97} = 1\} \\ &= E\{T_{02} \cdot Y_{02}(1,1)\} + E\{(1 - T_{02}) \cdot Y_{02}(1,0)\} \\ &= p_{02} \cdot E\{Y_{02}(1,1)\} + (1 - p_{02}) \cdot E\{Y_{02}(1,0)\} \end{aligned}$$

[22] See Heckman, Ichimura, Smith, and Todd (1998) for a formal proof that the lack of common support introduces bias.

and

$$\begin{aligned} E\{Y_{02} \mid T_{97}=0\} &= E\{T_{02}\cdot Y_{02}(0,1) \mid T_{97}=0\} \\ &\quad + E\{(1-T_{02})\cdot Y_{02}(0,0) \mid T_{97}=0\} \\ &= E\{T_{02}\cdot Y_{02}(0,1)\} + E\{(1-T_{02})\cdot Y_{02}(0,0)\} \\ &= p_{02}\cdot E\{Y_{02}(0,1)\} + (1-p_{02})\cdot E\{Y_{02}(0,0)\} \end{aligned}$$

so that

$$\begin{aligned} &E(Y_{02} \mid T_{97}=1) - E(Y_{02} \mid T_{97}=0) \\ &= p_{02}\left(E\{Y_{02}(1,1)\} - E\{Y_{02}(0,1)\}\right) \\ &\quad + (1-p_{02})\left(E\{Y_{02}(1,0)\} - E\{Y_{02}(0,0)\}\right). \end{aligned}$$

In our hypothetical example, this implies that

$$E(Y_{02} \mid T_{97}=1) - E(Y_{02} \mid T_{97}=0) = (1-p_{02})\widetilde{\tau} + (1-p_{02})K, \quad \textbf{(15)}$$

with a bias equal to $K - p_{02}(K+\widetilde{\tau})$. As seen earlier, when $\widetilde{\tau}=0$ and we condition on $T_{02}=0$, the bias is $K$. Because $K>0$, it follows from (15) that when $\widetilde{\tau}=0$ the bias in the entire population is $(1-p_{02})K < K$, which means that using the subpopulation for which $T_{02}=0$ leads to *higher* bias than using the entire population when the true treatment effect in the absence of the second randomization is 0. More generally, in this example the bias in the entire population will be less than in the subpopulation with $T_{02}=0$ for all $\widetilde{\tau}$ such that $-K < \widetilde{\tau} < K(2-p_{02})/p_{02}$.

## DISCONTINUITY IN VOTE SHARES AND FRESHMEN INCUMBENCY ADVANTAGE

We now provide some formal derivations for the discussion presented in the section *A Natural Experiment Based on a Discontinuity*. We show that the average of several different RD effects is in general different from the pooled effect for the case of two elections, and that having equal treatment and control sample sizes inside a window around the discontinuity in both elections is a sufficient condition for their equality. The vote shares in election 1 in bare-winner and bare-loser districts are denoted $y_1^W$ and $y_1^L$, respectively, and the number of bare-winner and of bare-loser observations are denoted $N_1^W$ and $N_1^L$, respectively. In general, these are observations on either side of the discontinuity within a given window around this discontinuity. The quantities $y_2^W$, $y_2^L$, $N_2^W$, and $N_2^L$ are defined analogously for election 2.

The effects in in elections 1 and 2 are, respectively, $\tau_1 = \frac{1}{N_1^W}\sum_{i=1}^{N_1^W} y_{i1}^W - \frac{1}{N_1^L}\sum_{i=1}^{N_1^L} y_{i1}^L$ and $\tau_2 = \frac{1}{N_2^W}\sum_{i=1}^{N_2^W} y_{i2}^W - \frac{1}{N_2^L}\sum_{i=1}^{N_2^L} y_{i2}^L$. The average effect is

$$\begin{aligned} \tau^{ave} = \frac{\tau_1+\tau_2}{2} &= \left(\frac{1}{2N_1^W}\sum_{i=1}^{N_1^W} y_{i1}^W + \frac{1}{2N_2^W}\sum_{i=1}^{N_2^W} y_{i2}^W\right) \\ &\quad - \left(\frac{1}{2N_1^L}\sum_{i=1}^{N_1^L} y_{i1}^L + \frac{1}{2N_2^L}\sum_{i=1}^{N_2^L} y_{i2}^L\right). \end{aligned}$$

The pooled effect is

$$\begin{aligned} \tau^{pool} &= \frac{1}{N_1^W+N_2^W}\left(\sum_{i=1}^{N_1^W} y_{i1}^W + \sum_{i=1}^{N_2^W} y_{i2}^W\right) \\ &\quad - \frac{1}{N_1^L+N_2^L}\left(\sum_{i=1}^{N_1^L} y_{i1}^L + \sum_{i=1}^{N_2^L} y_{i2}^L\right). \end{aligned}$$

In general, $\tau^{ave} \neq \tau^{pool}$. But if $N_1^W = N_2^W = N^W$ and $N_1^L = N_2^L = N^L$ we have

$$\begin{aligned} \tau^{pool} &= \frac{1}{2N^W}\left(\sum_{i=1}^{N^W} y_{i1}^W + \sum_{i=1}^{N^W} y_{i2}^W\right) \\ &\quad - \frac{1}{2N^L}\left(\sum_{i=1}^{N^L} y_{i1}^L + \sum_{i=1}^{N^L} y_{i2}^L\right) = \tau^{ave}. \end{aligned}$$

## REFERENCES


Abrajano, Marisa A., Jonathan Nagler, and R. Michael Alvarez. 2005. "A Natural Experiment of Race-Based and Issue Voting: The 2001 City of Los Angeles Elections." *Political Research Quarterly* 58 (2): 203–18.

Ansolabehere, Stephen, James M. Snyder, and Charles Stewart. 2000. "Old Voters, New Voters, and the Personal Vote: Using Redistricting to Measure the Incumbency Advantage." *American Journal of Political Science* 44 (1): 17–34.

Bhavnani, Rikhil R. 2009. "Do Electoral Quotas Work after They Are Withdrawn? Evidence from a Natural Experiment in India." *American Political Science Review* 103 (1): 23–35.

Butler, Daniel Mark. 2009. "A Regression Discontinuity Design Analysis of the Incumbency Advantage and Tenure in the U.S. House." *Electoral Studies* 28 (2): 123–28.

Carman, Christopher, James Mitchell, and Robert Johns. 2008. "The Unfortunate Natural Experiment in Ballot Design: The Scottish Parliamentary Elections of 2007." *Electoral Studies* 27 (3): 442–59.

Carson, Jamie L., Erik J. Engstrom, and Jason M. Roberts. 2007. "Candidate Quality, the Personal Vote, and the Incumbency Advantage in Congress." *American Political Science Review* 101 (2): 289–301.

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.

Cox, Gary W., and Jonathan N. Katz. 2002. *Elbridge Gerry's Salamander: The Electoral Consequences of the Reapportionment Revolution*. New York: Cambridge University Press.

Desposato, Scott W., and John R. Petrocik. 2003. "The Variable Incumbency Advantage: New Voters, Redistricting, and the Personal Vote." *American Journal of Political Science* 47 (1): 18–32.

Diamond, Alexis, and Jasjeet S. Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." Working Paper. http://sekhon.berkeley.edu/papers/GenMatch.pdf (accessed November 16, 2011).

Diamond, Jared, and James A. Robinson, eds. 2010. *Natural Experiments of History*. Cambridge, MA: Belknap Press.

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Science Quarterly* 61 (2): 282–93.

Erikson, Robert S. 1971. "The Advantage of Incumbency in Congressional Elections." *Polity* 3 (3): 395–405.

Gelman, Andrew, and Gary King. 1990. "Estimating Incumbency Advantage without Bias." *American Journal of Political Science* 34 (4): 1142–64.

Gordon, Sandy, and Greg Huber. 2007. "The Effect of Electoral Competitiveness on Incumbent Behavior." *Quarterly Journal of Political Science* 2 (2): 107–38.

Grofman, Bernard, Robert Griffin, and Gregory Berry. 1995. "House Members Who Become Senators: Learning from a 'Natural Experiment' in Representation." *Legislative Studies Quarterly* 20 (4): 513–29.

Hahn, Jinyong, Petra Todd, and Wilbert van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69: 201–09.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66 (5): 1017–98.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65 (2): 261–94.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Kaushik, Susheela. 1992. "Women and Political Participation." In *Women in Politics: Forms and Processes*, ed. Kamala Sankaran. New Delhi: Friedrich Ebert Stiftung.

Kishwar, Madhu. 1996. "Women and Politics: Beyond Quotas." *Economic and Political Weekly* 31 (43): 2867–74.

Krasno, Jonathan S., and Donald P. Green. 2008. "Do Televised Presidential Ads Increase Voter Turnout? Evidence from a Natural Experiment." *Journal of Politics* 70 (1): 245–61.

Lassen, David D. 2005. "The Effect of Information on Voter Turnout: Evidence from a Natural Experiment." *American Journal of Political Science* 49 (1): 103–18.

Lee, David S. 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics* 142 (2): 675–97.

Neyman, Jerzy. [1923] 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5 (4): 465–72. Trans. Dorota M. Dabrowska and Terence P. Speed.

Posner, Daniel N. 2004. "The Political Salience of Cultural Difference: Why Chewas and Tumbukas Are Allies in Zambia and Adversaries in Malawi." *American Political Science Review* 98 (4): 529–45.

Rosenbaum, Paul R. 2004. "Design Sensitivity in Observational Studies." *Biometrika* 91 (1): 153–64.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. New York: Springer-Verlag.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (6): 688–701.

Sekhon, Jasjeet S. 2008. "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods." In *The Oxford Handbook of Political Methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. New York: Oxford University Press, 271–99.

Sekhon, Jasjeet S. 2010. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12: 487–508.

Sekhon, Jasjeet S. 2011. "Matching: Multivariate and Propensity Score Matching with Automated Balance Search." *Journal of Statistical Software* 42 (7): 1–52. http://sekhon.berkeley.edu/matching/ (accessed November 16, 2011).

Sekhon, Jasjeet S., and Richard Grieve. N.d. "A Non-parametric Matching Method for Bias Adjustment with Applications to Economic Evaluations." *Health Economics*. Forthcoming.

Singh, Gopal Simrita, Medha Kotwal Lele, Nirmala Sathe, Wandana Sonalkar, and Anjali Maydeo. 1992. "Participation of Women in Electoral Politics in Maharashtra." In *Women in Politics: Forms and Processes*, ed. Kamala Sankaran. New Delhi: Friedrich Ebert Stiftung, 63–108.

van der Brug, Wouter. 2001. "Perceptions, Opinions and Party Preferences in the Face of a Real World Event: Chernobyl as a Natural Experiment in Political Psychology." *Journal of Theoretical Politics* 13 (1): 53–80.

Whitford, Andrew B. 2002. "Decentralization and Political Control of the Bureaucracy." *Journal of Theoretical Politics* 14 (2): 167–93.