Patterns of Effects and Sensitivity Analysis for Differences-in-Differences

Luke Keele* Dylan S. Small† Jesse Y Hsu‡

June 29, 2016

Abstract

In the estimation of causal effects with observational data, applied analysts often use the differences-in-differences (DID) method. The method is widely used since the needed before and after comparison of a treated and control group is a common situation in the social sciences. Researchers use this method since it protects against a specific form of unobserved confounding. Here, we develop a set of tools to allow analysts to better utilize the method of DID. First, we articulate the hypothetical experiment that DID seeks to replicate. Next, we outline the form of matching that allows for covariate adjustment for the DID method that is consistent with the hypothetical experiment. We also summarize a set of confirmatory tests that should hold if DID is a valid identification strategy. Finally, we adapt a well known method of sensitivity analysis for hidden confounding to the DID method. We develop these sensitivity analysis methods for both binary and continuous outcomes. We then apply our methods to two different empirical examples from the social sciences.

*Associate Professor, Department of Political Science, 211 Pond Lab, Penn State University, University Park, PA 16802 Email: ljk20@psu.edu, corresponding author.

†Professor, Department of Statistics, 400 Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104. E-mail: dsmall@wharton.upenn.edu

‡Assistant Professor of Biostatistics, Perelman School of Medicine, 423 Guardian Dr, Philadelphia, PA 19104, Email: hsu9@mail.med.upenn.edu

1 Introduction

1.1 Casual Inference and Public Policy

The need to understand the relationship between cause and effect is an essential part of public policy. Effective policymaking requires understanding the causal effects of proposals in order to devise the optimal policy. The need to understand relationships between cause and effect arises in almost every policy domain, including health, labor, education, environmental studies, public safety, and national security.

It well understood that randomized policy evaluations are the "gold-standard," since randomization ensures that subjects are similar accept for receipt of the treatment of interest. However, many policy evaluations occur in settings where randomized experiments are difficult or impossible. When randomized interventions are not possible, researchers may conduct an observational study. Cochran and Chambers (1965) defined an observational study as an empirical analysis where the objective is to elucidate cause-and-effect relationships in contexts where subjects select their own treatment status. When subjects select into treatments, outcomes may reflect pretreatment differences in treated and control groups rather than treatment effects (Cochran and Chambers 1965; Rubin 1974). Pretreatment differences in treated and control groups arise for either measurable differences which form overt biases or unmeasured differences which form hidden biases. In an observational study, analysts use pretreatment covariates and a statistical adjustment strategy such as matching or regression modeling to remove overt biases in the hopes of consistently estimating treatment effects.

It is also well understood, however, that such statistical adjustments do little to ensure that estimated treatment effects do not reflect hidden bias from confounders that were not included in the statistical adjustments. As such, investigators often employ devices, which consist of information collected in hopes of distinguishing an estimated association from bias (Rosenbaum 2010). One such device is the method of differences-in-differences. Differences-in-differences (DID) is used to distinguish an estimated treatment effect from bias by studying

a single treatment using four different groups where only certain patterns of response among the four groups are compatible with a treatment effect. In the simplest DID design, the analysts observes treated and control groups before and after the treatment is administered.

The method of DID is used to evaluate treatments across a wide range of policy domains. One famous example based on a DID design studied the effect Mariel Boatlift from Cuba on employment rates in the Miami labor market Card (1990). Another well known example based on differences-in-differences is in Dynarski (1999). Here, she studies the treatment effect of the additional aid on the decision to attend college, using changes in the Social Security Student Benefit Program, which awarded college aid to high school seniors with deceased fathers of Social Security recipients. Card and Krueger (1994) use DID to study changes in minimum wage laws on levels of employment. Hanmer (2009) uses DID to study whether changes voter registration laws increase voter turnout.

While the DID method does protect against a specific form of unobserved bias, it may still be the case that subjects differ with respect to an unmeasured covariate. Given uncertainty about the possibility of bias from unmeasured covariates, it is often useful to conduct a sensitivity analysis. A sensitivity analysis asks how strong the effects of an unmeasured covariate would have to be to substantively alter the conclusions from the study. In this study, we outline a method of sensitivity analysis for differences-in-differences. In addition, we describe a specific testing plan, which better allows analysts to judge whether a design based on differences-in-differences is plausible. This testing plan is based on the implied experiment that underlies a differences-in-differences design. We show that while an observational study based on differences-in-differences has some advantages, the method of differences-in-differences in many ways offers little protection against bias from hidden confounders, and its use would benefit from attention to some oft ignored points.

In this essay, we first explore the expected pattern of effects in a DID design. This structure for the effects in a DID analysis, better allows analysts to reason about the validity of the identification assumptions hold. We then outline the implied randomized experiment

that a DID analysis mimics. While covariate adjustment is common in DID applications, such adjustments are based on regression models, which impose strong functional form assumptions. We articulate a covariate adjustment strategy based on matching that mimics the implied experiment. We argue that covariate adjustment based on matching more closely follows the implied experiment and can reveal important differences between the treated and control group that may be missed if regression models are used. We then develop methods for sensitivity analysis. Our method for sensitivity analysis directly build on method of sensitivity analysis outlined in Rosenbaum (2002). We show that in several important ways, DID designs are quite sensitive to bias from hidden confounders. Finally, we conclude with two different empirical applications. In each application, we draw important lessons about how to judge whether an analysis based on DID is likely subject to bias from hidden confounders.

2 The Method of Differences-in-Differences

Observational studies that adopt a DID design share a common structure where a longitudinal component is observed along with an instance where a nonrandomly assigned treatment is applied to one group but not another. In each case, outcomes are observed for both the treated and control group before the treated group receives the treatment. Outcomes are then observed after the treatment has been administered to the treatment group. For example, in one of the applications below, we study whether the ability to register to vote on election day increases turnout. Specifically, we study when Wisconsin adopted election day registration in 1976. Here, residents of Wisconsin form the treated group, and we could designate residents of any state that did not adopt EDR as the control group. Turnout rates are observed in both the treated and control group before and after adoption of EDR. The DID estimate of the EDR treatment effect is based on two steps. First, the analyst take the difference in turnout rates before EDR adoption and the difference in turnout rates after the adoption of EDR. Second, the over time difference in these two group differences is taken.

DID can be applied in cases where the same set of subjects are observed in the before and after period, but it can also be applied if different sets of subjects are observed in each time period.

DID produces valid treatment effect estimates so long as any confounders are time invariant. That is, any confounders that affect the probability of turnout across treated and control group must be fixed across the before and after periods. Alternatively, we must assume that no other events beside the treatment alters the temporal path of either the treated or control groups. Next, we outline how a pattern of effects speaks to whether a DID design is likely to yield valid treatment effect estimates.

2.1 Pattern of Effects for DID

One strength of DID is that it lends itself to the use of pattern specificity. Statistical results from a single observational study are rarely considered to provide definitive proof of a causal relationship. In observational studies, where randomization is not present, analysts should acknowledge greater uncertainty about the effects caused by treatments than would have been present had treatments been randomly assigned. As such, a pattern of specific confirmatory tests provides better evidence than a single test. As Cook and Shadish (1994, pg. 95) write: "Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations." Designs based on DID lend themselves to a series of tests about patterns in the data that should hold if the design is valid.

In "Reforms as Experiments," Campbell (1969) discussed studies of the effects of institutional reforms. In particular, he discussed studies that measure institutions before and after the reform, as well measuring unreformed control institutions at parallel times. He offered an insightful discussion on the barriers to the success of such studies. Following his discussion, Figure 1 illustrates several issues in studies of this kind. Figure 1 depicts the median outcome in treated and control groups, in the periods before and after treatment in the treated group. The lower portion of Figure 1 depicts the corresponding situations

after a log transformation of the outcome. The log transformation is intended to be just one representative of the family of strictly increasing transformations.

Among the examples in Figure 1, case A is the most convincing: treated and control groups had similar outcomes prior to treatment, the control group did not change, but the outcomes increased in the treated group, and of course the log transformation changes the magnitudes but not the pattern. In case A in Figure 1, three different quantities all suggest the same effect of the treatment at the median: the post-treatment difference between treated and control groups, the change from base-line in the treated group, and the interaction or difference-in-differences. In the absence of random assignment of subjects to time periods and treatment conditions, each of these three comparisons is potentially biased, but certain biases that could affect one comparison are less likely to affect another, so each of the three comparisons provides a limited but useful check on the others.

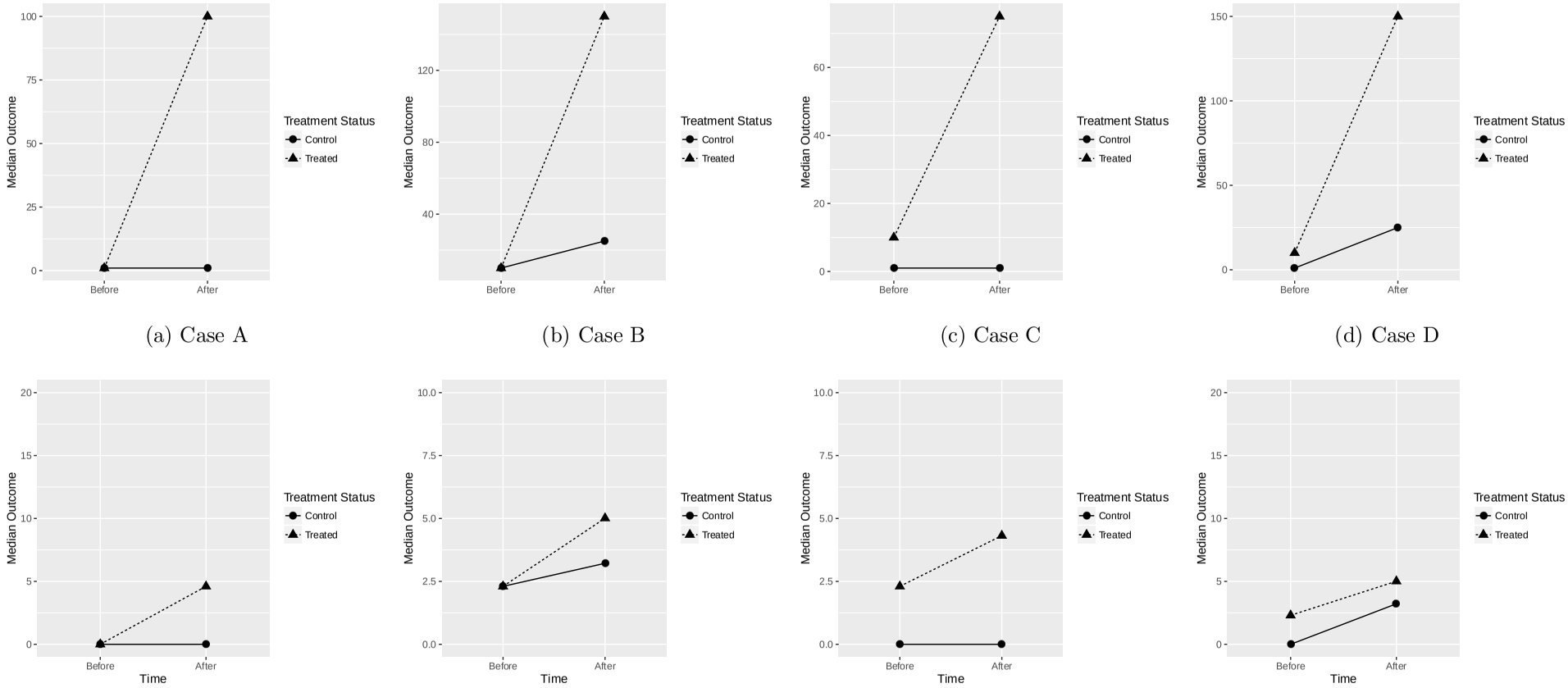
Case B is less convincing but not totally unconvincing: treated and control groups had similar outcomes prior to treatment and very different outcomes after treatment, but the control group changed in the absence of treatment, and of course the log transformation changes the magnitudes but not the pattern. In case B, the change from baseline in the treated group is not a plausible estimate because the controls also changed, but the post-treatment difference and the interaction produce the same estimate of effect.

Case C is also less convincing than case A, and arguably less convincing than case B: the groups were not comparable prior to treatment, but the treated group changed while the control group did not, and the log transformation changes magnitudes but not the pattern. In case C, the post-treatment difference is not a plausible estimate of effect, but the change in the treated group and the interaction produce the same estimate of effect.

Case D is the least convincing, perhaps totally unconvincing: the groups were not comparable prior to treatment, both groups changed, but the treated group changed by a larger amount; however, the changes look the same on the log scale. Although one might consider the interaction as an estimate of effect in case D, Campbell cautioned against a casual as-

sumption that the interaction estimates the effect. In particular, it would be at least very odd to say — arguably quite incorrect to say — that a treatment has an effect on an outcome but not on the log of that outcome, so there is something very odd about the interaction as an estimate of effect in case D in Figure 1. Even in the most convincing case, case A, an additional pre-treament measure one period before the plotted pretreatment measure might reveal a lazy X pattern with the cross at the shared before point, so that both groups were on a linear trajectory that did not change after treatment, suggesting no treatment effect.

Unlike Campbell, some investigators act as if the interaction or difference-in-differences is necessarily unbiased or less biased than the other comparisons, but it is difficult to imagine a rationale, beyond wishful thinking, that would support that view on a general basis. One could easily imagine a selection bias that is confounded with both selection into treatment and the treatment–time interaction. Following Campbell, we regard the entire pattern of responses in the four groups as relevant to evaluating the strength of evidence about an effect of the treatment and about possible biases. It seems to us that the strongest evidence would show a change from before reform to after in the treated group, a difference between treated and control after reform, a larger change from before to after in the treated than in the control group, and that the control group did not greatly change from before to after or at least followed the same trajectory as the treated group. What sort of analysis would best evaluate the evidence about effects and unobserved biases in the studies depicted in Figure 1? While we advocate visual inspection of plots like those in Figure 1, Figure 1 depicts unambiguous cases: groups are either identical or sharply different. In practice, with data, we may need to address marginal or ambiguous cases: groups that differ perhaps only by chance, groups that differ only slightly but by more than chance. We next develop a set of method for evaluating the evidence about effects and unobserved biases for the DID method.



(a) Case A (b) Case B (c) Case C (d) Case D

(e) Case A, Log-Scale (f) Case B, Log-Scale (g) Case C, Log-Scale (h) Case D, Log-Scale

Figure 1: chematic representation of response in treated and control groups, before and after treatment, with and without transformation to log scale.

2.2 DID: The Implied Experiment

Next, we develop formal notation for the DID method. We develop notation based on the experimental design that would produce the pattern of effects implied under a DID design, since one approach to the planning and design of observational studies is to study the similarities to and differences from an analogous randomized experiment (Cochran and Chambers 1965; Rubin 1974). We later use matching as the method to adjust for overt bias, as such, the implied experiment and the accompanying notation reflects the pairing produced by the matching process.

There are I matched sets, $i = 1, \ldots, I$, where each set i contains 4 subjects, $j = 1, 2, 3, 4$, with each subject assigned to one of 4 distinct conditions a, b, c, and d. Subjects in condition a are assigned to the treatment group and their outcomes are recorded after the treatment is applied, and subjects in condition b are assigned to the treatment group before the treatment is applied. Next, subjects in c are control subjects with outcomes recorded after treatment is applied, and subjects in condition d are control subjects with outcomes recorded before the treatment is administered. The I sets are matched for observed covariates $\mathbf{x}$, so that $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{x}_{i3} = \mathbf{x}_{i4}$ for all i.

Under a randomized design, if the jth subject in matched set i is assigned to group $k \in \{a, b, c, d\}$, write $Z_{ij} = k$. Then $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})$ is a permutation of $\{a, b, c, d\}$ for each i. Let $\mathcal{K} = \{abcd, abdc, \ldots, dcba\}$ be the set containing the $4! = 24$ possible values of $\mathbf{Z}_i$ formed by permuting the letters a, b, c, d. Let $\mathbf{Z}$ be the matrix with I rows and 4 columns whose I rows are the $\mathbf{Z}_i$, and let $\mathcal{Z}$ be the set containing the $(4!)^I$ possible values $\mathbf{z}$ of $\mathbf{Z}$, so $\mathbf{z} \in \mathcal{Z}$ if each row of $\mathbf{z}$ is a permutation of $\{a, b, c, d\}$. Also, denote the cardinality of a finite set $\mathcal{S}$ by $|\mathcal{S}|$, so $|\mathcal{K}| = 4!$ and $|\mathcal{Z}| = (4!)^I$. A randomized block experiment would use random numbers to pick a $\mathbf{z}$ at random, each $\mathbf{z} \in \mathcal{Z}$ having the same probability $|\mathcal{Z}|^{-1} = (4!)^{-I}$. This design enforces $\mathbf{Z} \in \mathcal{Z}$. For brevity, with a slight abuse of notation, conditioning on the event $\mathbf{Z} \in \mathcal{Z}$ is abbreviated as conditioning on $\mathcal{Z}$. Such an experiment would randomly assigns units to treatment or control in two specific time periods.

Each subject ij has a potential outcome under each condition $k \in \{a, b, c, d\}$, so ij would exhibit response r_{ijk} if ij received treatment k with $Z_{ij} = k$, but because each subject is seen under only one treatment, treatment effects such as $r_{ija} - r_{ijc}$ are not observed for any subject ij; see Neyman (1923) and Rubin (1974). The response actually observed from ij is R_{ij} which equals r_{ijk} if $Z_{ij} = k \in \{a, b, c, d\}$. Also, write Y_{ik} for the response R_{ij} of the subject in block i who received treatment k, that is, the subject with $Z_{ij} = k$; then, $Y_{ia} - Y_{ib}$ is the before-after change in the treated group, and $(Y_{ia} - Y_{ib}) - (Y_{ic} - Y_{id})$ is the interaction or difference-in-difference contrast. Fisher's (1935) sharp null hypothesis H_0 of no effect of any kind asserts $r_{ija} = r_{ijb} = r_{ijc} = r_{ijd}$ for all subjects ij.

Each subject ij has observed covariates $\mathbf{x}_{ij}$ and an unobserved covariate u_{ij}, and sets were matched for $\mathbf{x}_{ij}$, so $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$ for all i, j, j', but after matching for $\mathbf{x}_{ij}$ subjects may differ in terms of u_{ij}, so possibly $u_{ij} \neq u_{ij'}$ for many or all i, j, j'. Write $\mathcal{F} = \{(r_{ij1}, \ldots, r_{ij4}, \mathbf{x}_{ij}, u_{ij}), i = 1, \ldots, I, j = 1, \ldots, J\}$. Write $\mathbf{u} = (u_{11}, u_{12}, \ldots, u_{IJ})^T$, $\mathbf{R} = (R_{11}, R_{12}, \ldots, R_{IJ})^T$ and $\mathbf{r}_k = (r_{11k}, r_{12}, \ldots, r_{IJk})^T$ for $k = 1, 2, 3, 4$. Below, we outline a method of matching so that $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{x}_{i3} = \mathbf{x}_{i4}$ for all i.

If the only aspect of the treatment condition $\{a, b, c, d\}$ that affected the response was the introduction of the treatment, then $r_{ijb} = r_{ijc} = r_{ijd}$ for all ij, and we refer to this as the hypothesis of an isolated effect of the treatment, consistent with case A in Figure 1. An isolated and additive effect τ of the treatment has $r_{ija} - \tau = r_{ijb} = r_{ijc} = r_{ijd}$ for all ij, also consistent with case A in Figure 1. Case B in Figure 1 is consistent with $r_{ijb} = r_{ijd}$ for all ij and case C is consistent with $r_{ijc} = r_{ijd}$ for all ij. Of course, if $r_{ijb} = r_{ijd}$ for all ij then also $\log(r_{ijb}) = \log(r_{ijd})$ for all ij, and this is consistent with logs changing the magnitudes but not the patterns in case B in Figure 1.

An experimental design of this type is relatively uncommon in practice. More typically, an observational study is based on this design where outcomes for treated and control groups are observed before and after a treatment is nonrandomly applied. Analysts then focus on the difference-in-difference contrast as the causal estimand of interest. Observational studies

based on the DID device are considered useful since, even when the treatment is self-selected, it protects against two specific forms of bias. The two specific forms of bias are a uniform time trend, which we denote λ_t, affecting both groups in the same way, and a constant difference between treated and control groups, which we denote λ_d, such that, if both distorting effects were present in addition to an additive treatment effect without other distorting effects, then $r_{ija} - \tau = r_{ijd} + \lambda_t + \lambda_d$, $r_{ijb} = r_{ijd} + \lambda_d$, and $r_{ijc} = r_{ijd} + \lambda_t$. We refer to this as the additive distortions model. When the additive distortions model is correct, the interaction contrast or difference-in-difference $(Y_{ia} - Y_{ib}) - (Y_{ic} - Y_{id})$ removes the additive biases λ_t and λ_d.

However, the protection against hidden bias offered by a DID design is mostly the result of arithmetic convenience rather than the plausibility that these two distorting effects are the sole source of bias. For example, consistent with case D in Figure 1, the additive distortions model might hold for $\log(r_{ijk})$ but not for r_{ijk}, or conversely, or it might hold for some other strictly increasing transformation of r_{ijk} but neither r_{ijk} nor $\log(r_{ijk})$. Therefore, the additive pattern of distorting effects comes and goes with strictly monotone transformations of the response, leading us to doubt that additivity can be the central issue in answering a question about treatment effects. The fact that a design based on DID offers a solution to this form of bias often appears to be the primary reason people assume the bias has this convenient form. As such, widespread use of DID appears to mostly result from this common pattern in data rather than a belief that these are the only two forms of bias.

2.3 Adjustment for Overt Bias Via Matching

As we noted above, use of the DID device protects against two distorting effects that might bias treatment effect estimates. However, statistical methods are often applied to correct for differences between treated and control groups on observed covariates. Next, we outline the form of matching needed to remove overt bias in the context of DID. Here, three different matches must be performed so that units are balanced both with respect to treatment and control arms, but also with respect to time period. First, we match treated

units to control units in the pretreatment time period. This removes possible differences across treated and control groups prior to treatment. Next, we match treated to control units in the post-treatment time period. After these first two matches, we now have two sets of matched pairs, one from the pretreatment time period and one set from the post-treatment time periods. Using these two sets of matched pairs, we next match pre-treatment pairs to post-treatment pairs. In sum, we match pairs from the pretreatment time period to pairs from the post-treatment time period based on observed covariates. This third match balances observed covariates with respect to time. The form of matching in each case is need not be specific. Ideally, the matching would be done using an optimization algorithm (Rosenbaum 1989; Ming and Rosenbaum 2000; Hansen 2004; Zubizarreta 2012). We implement the matching in the application below using a method based on integer programming (Zubizarreta 2012). This form of matching allows us to specify specific balance constraints for each covariate. We implemented the matches using the `R` package `designmatch` (Zubizarreta and Kilcioglu 2016). We advocate matching as the method of adjustment since it allows us to remove overt biases without reference to outcomes. This prevents explorations of the data that may invalidate inferential methods (Rubin 2007). Moreover, matching does not impose restrictive functional form constraints required for more conventional methods of adjustment based on regression modeling.

As a practical matter, we match matched pairs to matched pairs across the time periods, we use summary statistics as covariates in the match. That is, for the matched pairs in each time period, we use the within pair mean as the covariate. For nominal covariates, the mean may not be a suitable summary within the pairs. In the applications that follow, we solve this problem by either exact matching or fine balance. Fine balance constrains an optimal match to exactly balance the marginal distributions of a nominal (or categorical) variable, perhaps one with many levels, placing no restrictions on who is matched to whom. This ensures that no category receives more controls than treated, and so the marginal distributions of the nominal variable are identical between the treatment and control groups. See Rosenbaum

et al. (2007) and Yang et al. (2012) for more details on fine balance. If we apply either fine balancing or exact matching to any nominal covariates in the initial matches, we can then exactly match or fine balance these covariates when we match pairs across the two time periods. The end result is matched sets such that $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{x}_{i3} = \mathbf{x}_{i4}$ for all i.

3 A Method of Sensitivity Analysis for DID

Next, we outline a method for sensitivity analysis that may be applied to DID estimates for the treatment effect. A sensitivity analysis allows an investigator to *quantify* the degree to which a key assumption must be violated in order for the original conclusion to be reversed. If an inference is sensitive, a slight violation of the assumption may lead to substantively different conclusions. We outline a sensitivity analysis method for DID based on a more general methods of bounds developed by Rosenbaum (2002). Under this method, one places bounds on quantities such as the treatment effect point estimate or p-value based on a conjectured level of confounding. We, first, outline the basic model for sensitivity analysis that we refer to as Rosenbaum bounds.

3.1 Model for sensitivity analysis: treatment assignments depend upon observed and unobserved covariates

In the population before matching, the unknown probability that subject ij is exposed to treatment k is

$$\pi_{ijk} = \Pr\left(Z_{ij} = k | \mathcal{F}\right) = \frac{\exp\left\{\xi_k\left(\mathbf{x}_{ij}\right) + \delta_k u_{ij}\right\}}{\sum_{\ell \in \{a,b,c,d\}} \exp\left\{\xi_\ell\left(\mathbf{x}_{ij}\right) + \delta_\ell u_{ij}\right\}}, \quad \mathbf{u} \in \mathcal{U}, \tag{1}$$

where $\mathcal{U} = [0,1]^{4I}$ is the $4I$-dimensional unit cube, $\xi_k(\cdot)$ is some unknown function, δ_k is an unknown sensitivity parameter, and treatment assignments for distinct subjects are independent. Under this model, the probability of assignment to treatment is solely a function

of observed covariates and u_{ij} an unobserved binary covariate. Write $\boldsymbol{\delta} = (\delta_a, \delta_b, \delta_c, \delta_d)^T$, and without loss of generality we may assume $\delta_k \geq 0$ for $k = a, b, c, d$, because replacing δ_k by $\delta_k - \min_{k' \in \{a,b,c,d\}} \delta_{k'}$ does not change π_{ijk} in (1). Model (1) says that two subjects, ij and $i'j'$, with the same observed covariate, $\mathbf{x}_{ij} = \mathbf{x}_{i'j'}$, may differ in their odds of receiving treatments k and k' by at most a factor of $\exp(\delta_k - \delta_{k'})$ because of $u_{ij} \neq u_{i'j'}$, that is,

$$\frac{1}{\exp(\delta_k - \delta_{k'})} \leq \frac{\pi_{ijk}(1 - \pi_{i'j'k'})}{\pi_{i'j'k'}(1 - \pi_{ijk})} \leq \exp(\delta_k - \delta_{k'}). \tag{2}$$

Generally, it is useful to have a single parameter Γ that summarizes the potential uncertainty due to the unknown vector $\boldsymbol{\delta}$, specifically: $\Gamma = \exp(\gamma)$ where $0 \leq \gamma = \max_{k \in \{a,b,c,d\}} \delta_k$, so $0 \leq \delta_k \leq \gamma$ for $k \in \{a, b, c, d\}$ and the odds ratio in (2) is at least $1/\Gamma = \exp(-\gamma)$ and at most $\Gamma = \exp(\gamma)$ for all k, $k' \in \{a, b, c, d\}$. In sum, if two subjects have the same observed covariates $\mathbf{x}$, then they may differ in their odds of receiving one of the four possible treatments by at most a factor of Γ. Two subjects, say ij and ij', with the same observed covariates, $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$, might be matched in the same block, and if $\Gamma = \exp(\gamma) = 1$ then these two subjects have the same unknown chance of receiving each treatment in (1), $\pi_{ijk} = \pi_{ij'k}$ for each k. However, if $\Gamma > 1$ then matching for $\mathbf{x}_{ij}$ failed to make the probability of treatment equal due to differences in u_{ij}.

If $\mathbf{k} \in \mathcal{K}$, then write $\boldsymbol{\delta}_{\mathbf{k}}$ for $(\delta_{k_1}, \delta_{k_2}, \delta_{k_3}, \delta_{k_4})^T$. For instance, with $\mathbf{k} = acbd$, $\boldsymbol{\delta}_{\mathbf{k}}$ is $(\delta_a, \delta_c, \delta_b, \delta_d)^T$. Matching for $\mathbf{x}_{ij}$ enforces $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$ and $\mathbf{Z}_i \in \mathcal{K}$ for all i, j, j'. Then conditioning on $\mathbf{Z}_i \in \mathcal{K}$ in (1), yields

$$\Pr(\mathbf{Z}_i = \mathbf{k} | \mathcal{F}, \mathbf{Z}_i \in \mathcal{K}) = \frac{\exp\left(\mathbf{u}_i^T \boldsymbol{\delta}_{\mathbf{k}}\right)}{\sum_{\mathbf{h} \in \mathcal{K}} \exp\left(\mathbf{u}_i^T \boldsymbol{\delta}_{\mathbf{h}}\right)}, \text{ for } \mathbf{k} \in \mathcal{K}, \tag{3}$$

so that $\Pr(\mathbf{Z}_i = \mathbf{k} | \mathcal{F}, \mathbf{Z}_i \in \mathcal{K}) = 1/|\mathcal{K}| = 1/24$ is the randomization distribution if $(\delta_a, \delta_b, \delta_c, \delta_d) = (0, 0, 0, 0)$, that is, if $\Gamma = 1$. A convenient feature of (3) is that, if $\mathcal{K}' \subseteq \mathcal{K}$, then

$$\Pr(\mathbf{Z}_i = \mathbf{k} | \mathcal{F}, \mathbf{Z}_i \in \mathcal{K}') = \frac{\exp\left(\mathbf{u}_i^T \boldsymbol{\delta}_{\mathbf{k}}\right)}{\sum_{\mathbf{h} \in \mathcal{K}'} \exp\left(\mathbf{u}_i^T \boldsymbol{\delta}_{\mathbf{h}}\right)}, \text{ for } \mathbf{k} \in \mathcal{K}'. \tag{4}$$

Expression (4) will supply for various $\mathcal{K}' \subseteq \mathcal{K}$ a sensitivity analysis for the comparison of treated and control groups before or after treatment as well as for the difference-in-differences of outcomes under the single model for treatment assignment in matched sets contained in (3).

3.2 Sensitivity analysis comparing two of the four groups

Suppose that we wish to compare two of the four groups, for example, after-treated b to after-control c. With the matching plan above, we have would have produced a form of matched pairs for this contrast. We could perform an outcome analysis using the I matched pair differences between the b and the c responses in the I blocks using Wilcoxon's signed rank statistic. With a suitable choice of $\mathcal{K}' \subset \mathcal{K}$, the conditional distribution in (4) reduces to a standard sensitivity analysis model for treated-minus-control matched pair differences.

If $\mathcal{K}' = \{\mathbf{k} \in \mathcal{K} : k_2 = c, k_4 = d\} = \{acbd, bcad\}$ is the set of $|\mathcal{K}'| = 2! = 2$ treatment assignments in which subject $j = 2$ received treatment c and subject $j' = 4$ received treatment d, then either subject 1 received b and subject 3 received c or else subject 1 received c and subject 3 received b, so $\mathcal{K}'$ contains the two permutations of b and c among subjects 1 and 3. In this case, (4) gives $\mathbf{k} = acbd$ conditional probability $\Pr(\mathbf{Z}_i = \mathbf{k} | \mathcal{F}, \mathbf{Z}_i \in \mathcal{K}')$ equal to

$$\Pr(\mathbf{Z}_i = acbd | \mathcal{F}, \mathbf{Z}_i \in \{acbd, bcad\}) = \frac{\exp(\delta_b u_{i1} + \delta_c u_{i3})}{\exp(\delta_b u_{i1} + \delta_c u_{i3}) + \exp(\delta_c u_{i1} + \delta_b u_{i3})}. \tag{5}$$

Because $0 \leq \delta_k \leq \gamma = \log(\Gamma)$ for $k \in \{a, b, c, d\}$ and $0 \leq u_{ij} \leq 1$, expression (5) is at most $\Gamma/(1+\Gamma)$ and is at least $1/(1+\Gamma)$. These bounds on (5) are sharp; for instance, the upper bound of $\Gamma/(1+\Gamma)$ is attained by $\delta_b = \gamma$, $\delta_c = 0$, $u_{i1} = 1$, $u_{i3} = 0$.

In general, when $\mathcal{K}' = \{\mathbf{k} \in \mathcal{K} : k_j = c, k_{j'} = d\}$, the sensitivity model (5) with bounds $\Gamma/(1+\Gamma)$ and $1/(1+\Gamma)$ is identical to the sensitivity analysis for a matched pair comparison of two treatments; e.g., Rosenbaum (1987, 2002). Therefore if we reduce the comparison to any two way comparison, treated–control or before–after, the form of sensitivity analysis

reduces to a standard application of Rosenbaum bounds. In the context of the matching plan we outline above, a sensitivity analysis may be performed for the two set of matched pairs (treated-control before treatment; treated-control after treatment) using standard methods. Next, we outline a sensitivity analysis for the differences-in-differences estimate.

3.3 Sensitivity analysis for the difference-in-differences

Next, we consider a sensitivity analysis of the DID treatment effect estimate. The DID treatment effect estimate sums the responses of the two subjects receiving conditions a and d (treatment and control in the pretreatment period) and subtracts the sum of the responses of the two subjects receiving conditions b and c, (treatment and control in the posttreatment period). For instance, if $\mathbf{Z}_i = dbca$ then the interaction contrast would be $(R_{i4} + R_{i1}) - (R_{i2} + R_{i3})$. Under the null hypothesis of no effect in a randomized experiment, the values $\pm\,|(R_{i4} + R_{i1}) - (R_{i2} + R_{i3})|$ would be equally probable as values of this contrast, leading to the conventional permutation distribution of, say, Wilcoxon's signed rank statistic. Next, we consider a sensitivity analysis that considers the possibility of biased treatment assignment, $\boldsymbol{\delta} \neq \mathbf{0}$ and how that might change our inference for the DID treatment effect estimate. To that end, we derive a sensitivity analysis using Expression (4).

Let $\mathcal{K}_1 \subset \mathcal{K}$ be the subset $\mathcal{K}_1 = \{abcd, acbd, dbca, dcba\}$ and let $\mathcal{K}_2 = \{badc, bdac, cadb, cdab\}$. If $\mathbf{Z}_i \in \mathcal{K}_1$ then the difference-in-difference contrast for set i would be $(R_{i1} + R_{i4}) - (R_{i2} + R_{i3})$, whereas if $\mathbf{Z}_i \in \mathcal{K}_2$ then the difference-in-difference contrast for set i would be $(R_{i2} + R_{i3}) - (R_{i1} + R_{i4})$. Conditioning on $\mathbf{Z}_i \in \mathcal{K}_1 \cup \mathcal{K}_2$ ensures the difference-in-difference contrast is either $(R_{i1} + R_{i4}) - (R_{i2} + R_{i3})$ or $(R_{i2} + R_{i3}) - (R_{i1} + R_{i4})$, under H_0, the absolute value of this contrast is fixed. Under (4), the conditional probability that $\mathbf{Z}_i \in \mathcal{K}_1$ given $\mathbf{Z}_i \in \mathcal{K}_1 \cup \mathcal{K}_2$ is

$$\Pr\left(\mathbf{Z}_i \in \mathcal{K}_1 \middle| \mathcal{F}, \mathbf{Z}_i \in \mathcal{K}_1 \cup \mathcal{K}_2\right) = \frac{\sum_{\mathbf{k} \in \mathcal{K}_1} \exp\left(\mathbf{u}_i^T \boldsymbol{\delta}_{\mathbf{k}}\right)}{\sum_{\mathbf{k} \in \mathcal{K}_1} \exp\left(\mathbf{u}_i^T \boldsymbol{\delta}_{\mathbf{k}}\right) + \sum_{\mathbf{h} \in \mathcal{K}_2} \exp\left(\mathbf{u}_i^T \boldsymbol{\delta}_{\mathbf{h}}\right)}. \tag{6}$$

If treatment assignment were randomized, then $\delta_a = \delta_b = \delta_c = \delta_d = \gamma = 0$ such that (6) equals 1/2. Next, we derive bounds on the probability of assignment contingent on the value of γ and u_{ij} a possible unobserved binary confounder.

Proposition 3.1 *If* $0 \leq \delta_k \leq \gamma$ *for each* $k \in \{a, b, c, d\}$ *and* $0 \leq u_{ij} \leq 1$ *for all* $j = 1, 2, 3, 4$, *then*

$$\frac{1}{1+\Gamma^2} \leq \Pr\left(\mathbf{Z}_i \in \mathcal{K}_1 \middle| \mathcal{F}, \mathbf{Z}_i \in \mathcal{K}_1 \cup \mathcal{K}_2\right) \leq \frac{\Gamma^2}{1+\Gamma^2}. \tag{7}$$

Moreover, the upper and lower bounds are sharp, being attained for particular u_{ij} *and* δ_k *with* $0 \leq u_{ij} \leq 1$ *and* $0 \leq \delta_k \leq \gamma$.

Proof. Of course, $\Gamma^2 = \exp(2\gamma)$. By algebra applied to (6), the inequality (7) is equivalent to

$$\exp(-2\gamma) \leq \frac{\sum_{\mathbf{k}\in\mathcal{K}_1} \exp\left(\mathbf{u}_i^T \boldsymbol{\delta}_\mathbf{k}\right)}{\sum_{\mathbf{h}\in\mathcal{K}_2} \exp\left(\mathbf{u}_i^T \boldsymbol{\delta}_\mathbf{h}\right)} \leq \exp(2\gamma). \tag{8}$$

The elements of $\mathcal{K}_1$ and $\mathcal{K}_2$ are in 1-to-1 correspondence: for each $\mathbf{k} = (k_1, k_2, k_3, k_4) \in \mathcal{K}_1$ there is a unique $\mathbf{k}' \in \mathcal{K}_2$ formed as $\mathbf{k}' = (k_2, k_1, k_4, k_3)$. Moreover, for $\mathbf{k}' \in \mathcal{K}_2$ corresponding to $\mathbf{k} \in \mathcal{K}_1$,

$$\mathbf{u}_i^T \boldsymbol{\delta}_\mathbf{k} - \mathbf{u}_i^T \boldsymbol{\delta}_{\mathbf{k}'} = (u_{i1} - u_{i2})(\delta_{k_1} - \delta_{k_2}) + (u_{i3} - u_{i4})(\delta_{k_3} - \delta_{k_4}). \tag{9}$$

Subject to $0 \leq \delta_k \leq \gamma$ for each $k \in \{a, b, c, d\}$ and $0 \leq u_{ij} \leq 1$, expression (9) is at most 2γ and at least -2γ. In other words, each term in the numerator of the ratio in (8) is at most $\exp(2\gamma)$ times greater than the corresponding term in the denominator, and each term in the numerator is at least $\exp(-2\gamma)$ times the corresponding term in the denominator, proving the inequality (8). If $u_{i1} = u_{i4} = 1$, $u_{i2} = u_{i3} = 0$, $\delta_a = \delta_d = \gamma$, and $\delta_b = \delta_c = 0$, then (9) equals 2γ for each $\mathbf{k} \in \mathcal{K}_1$ and its corresponding $\mathbf{k}' \in \mathcal{K}_2$, thereby achieving the upper bound in (8) and hence in (7). The lower bound is analogous. ■

The results of this proof have two important implications. First, this result means that to conduct a sensitivity analysis for the DID treatment effect estimate, an investigator can use a sensitivity analysis for matched pairs applied to the difference-in-difference contrast– $(Y_{ia} - Y_{ib}) - (Y_{ic} - Y_{id})$, but with Γ^2 replacing Γ. Second, we observe that the DID treatment effect estimate is actually *more* sensitive to hidden bias due to an unobserved confounder. The reason this is true is that the same magnitude of bias Γ in assigning any two treatments can have a larger effect Γ^2 on the interaction contrast because the interaction contrast is affected by four rather than two treatment assignments. That is, a hidden confounder might alter the probability of being assigned to treatment or control, but also the before and after time period. For example, a bias of the form $\delta_a = \delta_d = \gamma$ and $\delta_b = \delta_c = 0$ could tilt higher responses towards the a and d conditions and away from the b and c conditions, strongly affecting the interaction contrast. Thus while the DID estimate protects against the additive distortions model, many other forms of bias may shift responses in many directions. Next, we derive one additional form of sensitivity analysis that is less conservative.

3.4 A sensitivity analysis assuming an estimable time trend

Next, we develop one additional form of sensitivity analysis based on assumptions about the time trend in the control group outcomes. Under this form of sensitivity analysis, we assume that we can consistently estimate the additive trend, λ_t from the before and after contrast in the control group. If we adjust the data by the estimated value for λ_t, we can assume to have eliminated the effect of the correlation between the unobserved confounder and changes in the outcome over time, and we can then apply a sensitivity analysis to the DID treatment effect estimate using Γ instead of Γ^2.

A sensitivity analysis of this form begs two questions. The first is how might we estimate λ_t? The second is whether it is reasonable to assume we have a consistent estimate for λ_t? The first question is a fairly mechanical one, and we outline methods for estimation below. The second is a more substantive question that depends on judgement, and it is

one that must be made by the investigator. As we detail below, we estimate λ_t from the over time changes in the control group. That is, we assume that temporal changes in the control group can be used to remove bias due to over time changes in the treatment group not attributable to the treatment itself. This would seems to be a reasonable when overall trends in the treated and control group appear comparable. This adjustment is similar in spirt to examining the treated and control groups for parallel trends in the pre-treatment time period. Next, we outline methods for estimating λ.

When the outcomes are binary this form of sensitivity analysis is relatively straightforward. As Zhang et al. (2012) show, a test developed by Gart (1969) for the analysis of matched proportions in a crossover design, can be directly applied to conduct a sensitivity analysis for the DID device when outcomes are binary. They demonstrate that by taking sets of discordant outcome pairs from the matched pairs in the pre-treatment period and the matched pairs in the post-treatment period, the extended hypergeometric distribution can be applied to test the sharp null that the DID treatment effect is zero. They also show that sensitivity bounds can be constructed using Γ^2 rather than Γ. This forms the general form of sensitivity analysis for the DID treatment effect with binary outcomes. However, Gart's test assumes a known time trend is zero, thus we can construct sensitivity bounds assuming a known time trend with the extended hypergeometric distribution and Γ.

When outcomes are not binary, we estimate λ_t using a method developed by Berger and Boos (1994). They propose allowing for a nuisance parameter in a hypothesis test by maximizing a p-value over values of the nuisance parameter, here λ_t, inside a $1-\beta$ confidence set for λ_t, and then increasing the resulting p-value by adding β. We use this method to obtain a confidence set for λ_t using $(Y_{ic} - Y_{id})$. Specifically, one computes a $1-\beta$ confidence set $\mathcal{C}$ for λ_t based on $(Y_{ic} - Y_{id})$, then tests $H_0 : \tau = \tau_0$ in the additive distortions model by testing the null hypothesis of no effect on $(Y_{ia} - Y_{ib}) - \tau_0 - \widetilde{\lambda}_t$ for every $\widetilde{\lambda}_t \in \mathcal{C}$, and increasing the maximum p-value by the addition of β. This produces a confidence interval for λ_t, which we denote this confidence as $[\lambda_{t-}, \lambda_{t+}]$.

For many test statistics, we can conduct a sensitivity analysis using these endpoints of the confidence interval for λ_t to adjust the data. For a given value of Γ, we then calculate two test statistics. The first test statistic, T_1, measures the standardized discrepancy based on $(Y_{ia} - Y_{ib}) - \lambda_{t+}$. The second test statistic, T_2, is based on the following contrast: $(Y_{ic} - Y_{id}) - \lambda_{t-}$. Thus we adjust the data by the smallest and largest plausible values for λ_t. The upper-bound on the upper-one-sided p-value for Γ is based the sum of β and the two-sided p-value from the minimum of the lower tail p-value based on T_2 and the upper tail p-value based on T_1. Therefore, if the investigator is willing to make assumptions about the time trend, a sensitivity analysis can be applied using Γ rather than Γ^2.

4 A Sensitivity Analysis Plan for Differences-in-Differences

In a DID design, one can apply a four different sensitivity analysis to three different treated and control contrasts. First, one can apply a sensitivity analysis to the treated and control contrast before treatment. Second one can apply a sensitivity analysis to the treated and control contrast after treatment. Third, one can apply a sensitivity analysis to the DID contrast, and finally one can apply a sensitivity analysis to the DID contrast assuming that the time trend can be estimated from the control group. The question we engage next is which of these sensitivity analyses should analysts use? Specifically, we outline plan for the application of sensitivity analysis in the context of a DID design. By plan, we mean a part of the design that outlines the specific forms of sensitivity analysis that will be applied before outcomes are considered.

We recommend the following analysis plan. First, analysts should conduct a sensitivity analysis for the DID contrast, without assumptions about time trends. This sensitivity analysis is the most conservative, but that conservatism directly arises from the fact that confounders may after effect either the treated and control contrast or from a shift in the temporal levels of the outcomes. Reporting any other sensitivity analysis denies the possi-

bility that hidden bias may take some more complex form that is assumed by a DID design. An analysis may choose to stop the sensitivity analysis at this point. Next, investigator may choose to report the sensitivity analysis that assumes the time trend is estimable. A critical point will be to then contrast whether there are clear differences between these two sensitivity analyses. If both demonstrate that our conclusions can be easily explained by a hidden confounder then the conclusions are consistent across both methods. If the assumption of a time trend renders the results less sensitive to hidden bias, then qualitative knowledge about the defensibility of estimating the trend form the control group should be presented.

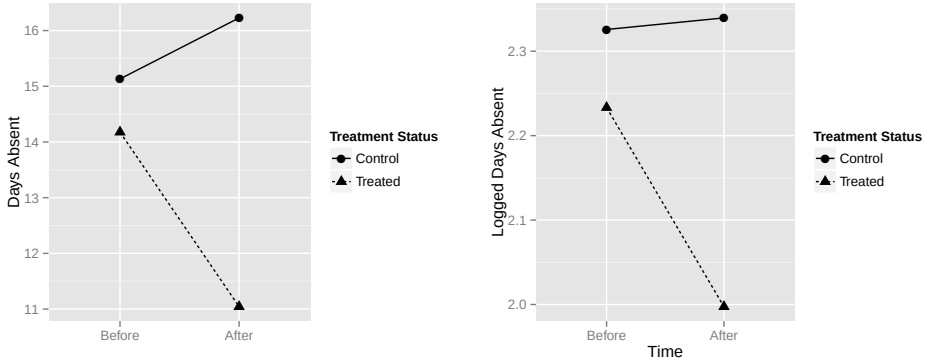
Finally, the analyst may ignore the temporal component of the study and report a sensitivity analysis for treated and control contrast in the post-treatment time period. This contrast makes no assumptions about time trends and after matching is a valid design that assumes treatment assignment is as-if random conditional on the observed covariates. Next, we review two different applications and demonstrate these methods in empirical settings.

5 Application: Disability Payments in Germany

In our second application, we re-analyze data from a study on whether a change in disability payment rates in Germany changed sick day usage Puhani and Sonderhof (2010). In 1995, Germany changed employment regulations such that workers who were covered by a collective bargaining contract (unionized workers) had their disability payments reduced from 100% coverage to 80% coverage. The goal in the original analysis was to understand whether the change in employment regulation contributed to workers using disability services at lower rates. The control group in the analysis is workers that are not covered by collective bargaining agreements. We focus on one of the outcomes from the original study: the number of days absent from work.

We begin the analysis with plots of the outcomes for the treated and control groups in both the before and after period. These figures are useful, since they allow us to observe

whether the unadjusted data fits the pattern of effects that should hold in a DID design. Figure 2 contains the basic results is graphical form, with both a logged and non-logged outcome. The pattern of effects in the data does not closely match what we expect under a DID design. In the treated group, we observe a clear decline in the number of days absent, but we also clearly observe a overtime change in the control group outcomes in the *opposite* direction. The graphical pattern here does not strongly recommend the sensitivity analysis that assumes the time trend is estimable from the control group given that the control group trend does not appear to mirror that in the treated group.



(a) Raw Outcomes (b) Log Scale

Figure 2: Outcomes for the German Disability Payments

We followed the matching plan we outlined above using cardinality matching (Zubizarreta et al. 2014). We implemented the match using the `R` package `designmatch`(Zubizarreta and Kilcioglu 2016). We match on the same set of covariates used in the original analysis. These include measures for hourly wage, age, education levels, blue or white collar status, firm size, length of tenure with company, and industry. For three nominal covariates, we set fine balance constraints. Under fine balance, we balance the marginal distribution of a categorical covariate so that it is exactly the same across the treated and control groups. (Rosenbaum et al. 2007). Fine balance constraints are not always feasible. One alternative is a near fine balance constraint which returns a finely balanced match when one is feasible, but minimizes the deviation from fine balance when fine balance is infeasible (Yang et al. 2012).

Specifically, we applied near fine balance constraints on firm size and length of tenure with your employer. We finely balanced industry. The appendix contains a table which reports within pair differences before and after matching. The matching resulted in 356 match pairs in the period before treatment went into effect.

Next, we applied the exact same form of matching to the treated and control units in the period after the change in disability payments. The match in the post-treatment period produced 470 matched pairs. Finally, we matched the 356 pairs from the pre-treatment period to the 470 matched pairs from the post-treatment period. To pair pairs to pairs, we took within pair averages within sets of matched pairs. For nominal covariates, we took the median with pairs. In this match, we exactly matched on industry and finely balanced both firm size and length of tenure with company. For the pair-to-pair match, we applied cardinality matching, which returns the largest set of matched pairs that met our pre-specified balance constraints (Zubizarreta et al. 2014). This resulted in 336 sets of pairs matched to pairs. Balance tables from both matches are also in the appendix.

Next, we estimated the DID treatment effect. The estimated DID treatment effect is -1.57, which implies that reducing disability payments reduced the average number of days absent from work by just under two days. This estimate relies on means to calculate the DID contrast, and the distribution for the number of days absent has a long tail. In case the tails of the distribution overly affect our estimate, we next use Wilcoxon's sign rank test to estimate the DID treatment effect. The estimate from the sign rank test is -2, with a one sided p-value of 0.0895, and a 95% CI of $(\infty, 0.50)$.

Finally, we also use an M-statistic with Huber's (1964; 1981) weight function. In an M-statistic, the observations are transformed to prevent a small number of observations from having a strong influence on the results. Moreover, an M-statistic is a continuous function of the integer days absent, so it is more graceful in handling the ties than the signed rank statistic. Results based on such M-estimates are often more resistant to hidden bias (Rosenbaum 2014). We implement M-estimates using functions from the `sensitivitymw`

package in R (Rosenbaum 2015a,b) with the defaults set for matched pairs. Using an M-statistic, we reject the sharp null with $p = 0.041$. The point estimate under M-estimation is -2.42, with a 95% confidence interval of $[-\infty, 0.13]$

Next, we conduct a sensitivity analysis for the DID treatment effect estimate. That is, we ask whether an unobserved confounder would have to change the odds of treatment by a small or large amount before our conclusions are reversed. We perform the sensitivity analysis for both the signed rank statistic and the M-statistic. As we outlined above, we an apply standard methods for Rosenbaum bounds using Γ^2 to calculate sensitivity at Γ. We begin by placing bounds on the one-sided p-value at $\Gamma = 1.01$. For the signed rank statistic, the upper-bound on the one-sided p-value is 0.13, and for the M-statistic the upper-bound is 0.05. Thus, in both cases, the estimate is extremely sensitive to bias from a hidden confounder. A hidden binary confounder would have to change the odds of treatment within matched pairs of pairs by a mere one percentage point.

We conduct an additional sensitivity analysis that assumes λ_t is estimable from the control group. As we noted above, this is a somewhat dubious assumption given that the control group moves in the opposite direction from the treated over time. However, we conduct the analysis to demonstrate how this assumption alters the testing. For this analysis, we only use M-statistics. First, we test the sharp null for the DID contrast when $\Gamma = 1$, and we can reject the sharp null $p = 0.02$. This result demonstrates why we recommend that this test not be done in isolation. In this application, reliance on a dubious assumption indicates a more decisive rejection of the sharp null. For $\Gamma = 1.07$, the upper-bound on the one-sided p-value is 0.045 if we assume the time trend is estimable. Thus, our conclusions are modestly more resistant to bias from a hidden confounder under this scenario. However, our conclusions here, rest on the assumption that the over time change in days absent can be estimated from the control group, which as we noted is not well justified in this setting.

Finally, we conclude with an analysis of the matched pairs after the treatment went into effect. That is, we can always simply use the standard methods for observational studies to

the treated-control contrast. This eliminates the need to assume the temporal component of the bias is additive. Using the sign rank test, we estimate that workers under a collective bargaining agreement were absent from work 1.5 more days than the comparable match controls. Here, we can reject the sharp null with $p = 0.04$. Using an M-estimate, the estimated treatment effect is -1.9 with a p-value of 0.017. For the signed rank, the upper-bound on the one-sided p-value is 0.06 when $Gamma = 1.02$. For the M-estimate, the upper-bound on the one-sided p-value is 0.054 when $Gamma = 1.08$. Again, in both cases, it would take very little to change our conclusions.

6 Application: Election Day Registration

The method of DID is often used to study the effect of policy changes in subnational units of government. For example many in the U.S. states allow voters to register to vote on election day. The goal behind election day registration (EDR) is to ease the burden of registration and allow more citizens vote. A number of studies have concluded that EDR has contributed to an increase in voter turnout. (Brians and Grofman 1999, 2001; Hanmer 2007, 2009; Highton and Wolfinger 1998; Knack 2001; Mitchell and Wlezien 1995; Rhine 1995; Teixeira 1992; Timpone 1998; Wolfinger and Rosenstone 1980). Though recent works suggest these studies are subject to substantial bias from hidden confounder (Keele and Minozzi 2012). As an illustration, we conduct a small scale study of EDR. In our application, we focus on Wisconsin one of the first states to adopt EDR and one place where the effect of EDR is widely understood to have contributed to an increase in turnout (Hanmer 2009).

Our data are based on extracts from the 1972 and 1980 Current Population Survey (CPS) and are based on a subset of data from a study by Keele and Minozzi (2012). The CPS is a monthly individual level survey conducted by the U.S. census which ask respondents about voting in the November survey of election years. Wisconsin first used EDR in 1976, we use turnout in the 1980 presidential election as the post-treatment period in case of any delay

in the effect of EDR. That is, it might take one election cycle before voters are fully aware that EDR is available. We use voters from Illinois as controls. Illinois would seem to be a reasonable counterfactual to Wisconsin. Illinois is adjacent to Wisconsin and both have large metropolitan areas with minority communities but also have large rural populations as well.

As before, we begin with a plot of the turnout rates in both states. Figure 3 contains the turnout before the implementation of EDR in Wisconsin and after for both states. The results are promising in that in 1972 both states had very similar levels of turnout. Moreover, we also observe a sharp increase in turnout in Wisconsin in 1980. However, the plot suggests that some factor or factors contributes to a sharp decrease in turnout in Illinois between 1972 and 1980.

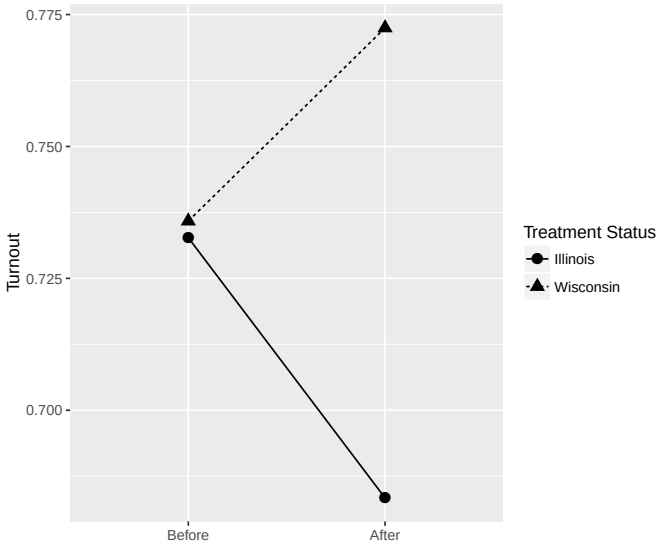


Figure 3: Outcomes for the EDR Example. Base Year is 1972, Outcome Year is 1980.

We begin the analysis by matching Wisconsin residents to Illinois residents, first in 1972, and then again in 1980. We match residents on age, an indicator if he or she is African American, female, a categorical scale for level of education, a categorical scale of income, and an interaction between education and income. In our match, we matched exactly on whether a resident was African-American, and we applied near fine balance to education,

income, and the interaction between education and income. We allowed for at maximum a deviation of two categories on the near fine balance in the match. After matching in 1972, we have 1427 matched pairs. After matching in 1980, we have 1718 matched pairs. We then performed the pair-to-pair match, where we matched the pairs from 1972 to pairs from 1980. Imbalances tended to much large across the two time periods that within each year. For the pair-to-pair match, we again applied cardinality matching. We are left with 790 matched pairs from 1972 matched to 790 matched pairs from 1980. Table 2 contains the results after the matching was completed. The upper part of Table 2 contains cross-tabulations of the outcomes for the matched pairs from 1972 and 1980. For each set of matched pairs, the table counts concordant and discordant outcomes—the number of discordant pairs are in the off-diagonal cells. To test Fisher's sharp null in either 1972 or 1980, we would apply McNemar's test individually to each of these two tables. Our interest, however, is in testing the sharp null for the DID treatment effect. To test, the sharp null for the DID treatment effect, we form a second table composed of the discordant pairs from the matches in 1972 and 1980. The lower part of Table 2 contains this new contingency table of discordant pairs. To this table, we apply Gart's test based on the extended hypergeometric distribution. Based on this test, the sharp null hypothesis is implausible ($p < 0.001$), though, of course, this test assumes there are no hidden confounders present.

Next, we seek to characterize the magnitude of the EDR effect. One simple method for summarizing the effect of EDR is to simply calculate the odds-ratio as applied to the lower table in Table 2. The estimated odds ratio is 1.79, which indicates that the presence of election day registration increase the odds of voting by 79%. Alternatively, we could simply calculate the DID treatment effect using proportions. According to to this estimate, the turnout rate increased 12.6 percentage points in Wisconsin as compared to Illinois. Compared to most interventions designed to increase voter turnout, this is a very large treatment effect. However, this estimate assumes that there is no bias from hidden confounders. Next, we turn to a sensitivity analysis to explore this possibility.

Table 1 contains the results from two different sensitivity analyses. The first makes no assumptions about λ_t, the nuisance time trend. Here, the sensitivity bound is calculated using the extended hypergeometric distribution using Γ^2. We find that the one-sided p-value is 0.05 when $\Gamma = 1.18$. This implies that an unobserved confounder could reverse our conclusions if it affect the odds of assignment to treatment or control in either time period by 18%. It is important to understand that this confounder could be correlated with a higher chance of being exposed to EDR or correlated with a change in the likelihood of voting over time. While not as sensitive to the possibility of bias from a confounder as the last example, this is a still a relatively small value for Γ. In the next sensitivity analysis, we assume that we can consistently eliminate bias from the effect of the confounder on over time changes in the likelihood of voting. In this instance, we again calculate the sensitivity analysis using the extended hypergeometric distribution, but we not use Γ instead of Γ^2. Now a moderate bias from u could produce this pattern of associations, $\Gamma = 1.39$.

Table 1: Sensitivity Analysis for the EDR Application. The table gives the upper bound on the one-sided p-value for the testing the null effect of EDR on voter turnout. Sens 1 assumes

Γ	Sensitivity Analysis with Unknown Trend	Sensitivity Analysis with Estimable Trend
1.00	0.00	0.00
1.05	0.00	0.00
1.10	0.01	0.00
1.15	0.03	0.00
1.18	0.05	0.00
1.25	0.20	0.01
1.30	0.37	0.02
1.39	0.71	0.05

7 Discussion

The method of DID is widely used to estimate causal effects. It is particularly useful when one region, state or country adopts a new public policy and other regions, states,

Table 2: Results for Differences-in-Differences for Election Day Registration in Wisconsin.

	1972		1980	
	Didn't Vote	Voted	Didn't Vote	Voted
Didn't Vote	51	161	98	150
Voted	159	567	266	424
		1980	1972	
Voted/Didn't Vote		266	159	
Didn't Vote/Voted		150	161	
Odds ratio		1.79		
p-value		7.14×10^{-5}		
95% Interval		$[1.32, 2.44]$		

and countries do not. Under this design, the hope is that the configuration of the bias from unobserved confounders has a specific additive form that can be eliminated when the investigator obtains data from treated and control groups before and after a treatment goes into effect. Of course, there are also reasons to think that a design based on DID offers little protection against bias. In general, we have no reason to think that the bias follows this particular configuration. This is reflected in the form of sensitivity analysis we have outline above. Since an observed confounder might shift the treated and control contrast or the trend in the treated group, the DID design is in a real sense doubly sensitive to bias from hidden confounders.

In addition to outlining methods for sensitivity analysis, we also outlined the hypothetical experiment on with the DID effect is based, as well as plan for covariate adjustment based on matching. Covariate adjustment based on matching makes much weaker functional form assumptions than the usual methods based on regression models.

In particular, DID does not depend on a known assignment rule. We know Wisconsin adopted EDR but we do not know if state legislators were seeking to maximize turnout or perhaps resolve a legislative compromise (Smolka 1977). A useful contrast is between DID design and a regression discontinuity (RD) design. In an RD design, a known treatment assignment rule is applied and respected. The strength of RD designs comes directly from

the use and application of this known assignment rule (Lee and Lemieux 2010).

References

Berger, R. L. and Boos, D. D. (1994), "P values maximized over a confidence set for the nuisance parameter," *Journal of the American Statistical Association*, 89, 1012–1016.

Brians, C. L. and Grofman, B. (1999), "When Registration Barriers Fall, Who Votes? An Empirical Test of a Rational Choice Model," *Public Choice*, 99, 161–176.

— (2001), "Election Day Registration's Effect on U.S. Voter Turnout," *Social Science Quarterly*, 82, 170–183.

Campbell, D. T. (1969), "Reforms as experiments." *American psychologist*, 24, 409.

Card, D. (1990), "The impact of the Mariel boatlift on the Miami labor market," *Industrial & Labor Relations Review*, 43, 245–257.

Card, D. and Krueger, A. B. (1994), "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *The American Economic Review*, 84, 772–793.

Cochran, W. G. and Chambers, S. P. (1965), "The Planning of Observational Studies of Human Populations," *Journal of Royal Statistical Society, Series A*, 128, 234–265.

Cook, T. and Shadish, W. (1994), "Social Experiments: Some Developments Over the Past Fifteen Years," *Annual Review of Psychology*, 45, 545–580.

Dynarski, S. M. (1999), "Does aid matter? Measuring the effect of student aid on college attendance and completion," *American Economic Review*, 93, 279–288.

Fisher, R. A. (1935), *The Design of Experiments*, London: Oliver and Boyd.

Gart, J. J. (1969), "An exact test for comparing matched proportions in crossover designs," *Biometrika*, 56, 75–80.

Hanmer, M. J. (2007), "An Alternative Approach to Estimating Who is Most Likely to Respond to Changes in Registration Laws," *Political Behavior*, 29, 1–30.

— (2009), *Discount Voting*, New York, NY: Cambridge University Press.

Hansen, B. B. (2004), "Full matching in an observational study of coaching for the SAT," *Journal of the American Statistical Association*, 99, 609–618.

Highton, B. and Wolfinger, R. E. (1998), "Estimating the Effects of the National Voter Registration Act of 1993," *Political Behavior*, 20, 79–104.

Huber, P. J. (1964), "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, 35, 73–101.

— (1981), *Robust Statistics*, New York, NY: John Wiley and Sons.

Keele, L. J. and Minozzi, W. (2012), "How Much is Minnesota Like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data," *Political Analysis*, 21, 193–216.

Knack, S. (2001), "Election-Day Registration: The Second Wave," *American Politics Research*, 29, 65–78.

Lee, D. S. and Lemieux, T. (2010), "Regression Discontiuity Designs in Economics," *Journal of Economic Literature*, 48, 281–355.

Ming, K. and Rosenbaum, P. R. (2000), "Substantial gains in bias reduction from matching with a variable number of controls," *Biometrics*, 56, 118–124.

Mitchell, G. E. and Wlezien, C. (1995), "Voter Registration and Election Laws in the United States, 1972-1992," *ICPSR*, 6496, 999.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science*, 5, 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).

Puhani, P. A. and Sonderhof, K. (2010), "The effects of a sick pay reform on absence and on health-related outcomes," *Journal of health economics*, 29, 285–302.

Rhine, S. (1995), "Registration Reform and Turnout Change in American States," *American Politics Quarterly*, 23, 409–427.

Rosenbaum, P. R. (1987), "Sensitivity Analysis For Certain Permutation Inferences in Matched Observational Studies," *Biometrika*, 74, 13–26.

— (1989), "Optimal Matching for Observational Studies," *Journal of the American Statistical Association*, 84, 1024–1032.

— (2002), *Observational Studies*, New York, NY: Springer, 2nd ed.

— (2010), *Design of Observational Studies*, New York: Springer-Verlag.

— (2014), "Weighted M-statistics with superior design sensitivity in matched observational studies with multiple controls," *Journal of the American Statistical Association*, 109, 1145–1158.

— (2015a), "`sensitivitymw`: Sensitivity analysis using weighted M-statistics," `R` package version 1.1.

— (2015b), "Two R packages for sensitivity analysis in observational studies," *Observ. Stud*, 1, 1–17.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007), "Mimimum Distance Matched Sampling with Fine Balance in an Observational Study of Treatmetnt for Ovarian Cancer," *Journal of the American Statistical Association*, 102, 75–83.

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

— (2007), "The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials," *Statistics in medicine*, 26, 20–36.

Smolka, R. G. (1977), *Election Day Registration: The Minnesota and Wisconsin Experience in 1976*, Washington, D.C.: American Enterprise Institute for Public Policy Research.

Teixeira, R. A. (1992), *The Disappearing American Voter*, Washington D.C.: Brookings.

Timpone, R. J. (1998), "Structure, Behavior, and Voter Turnout in the United States," *American Political Science Review*, 92, 145–158.

Wolfinger, R. E. and Rosenstone, S. J. (1980), *Who Votes?*, New Haven: Yale University Press.

Yang, D., Small, D. S., Silber, J. H., and Rosenbaum, P. R. (2012), "Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes," *Biometrics*, 68, 628–636.

Zhang, K., Traskin, M., and Small, D. S. (2012), "A Powerful and Robust Test Statistic for Randomization Inference in Group-Randomized Trials with Matched Pairs of Groups," *Biometrics*, 68, 75–84.

Zubizarreta, J. R. (2012), "Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure after Surgery," *Journal of the American Statistical Association*, 107, 1360–1371.

Zubizarreta, J. R. and Kilcioglu, C. (2016), "`designmatch`: Construction of Optimally Matched Samples for Randomized Experiments and Observational Studies that are Balanced by Design," `R` package version 0.1.1.

Zubizarreta, J. R., Paredes, R. D., and Rosenbaum, P. R. (2014), "Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile," *The Annals of Applied Statistics*, 8, 204–231.

Appendices

A.1 Balance Tables for First Application

Table 3: Standardized Differences and p-values for Treated to Control Match in the Pre-treatment Period for the Disability Payments Application

	Before Matching		After Matching	
	Std Dif	P-val	Std Dif	P-val
Regional Unemp.	0.19	0.00	-0.00	0.97
Hourly Wage	-0.04	0.51	0.07	0.39
Age	-0.10	0.09	-0.03	0.71
Married	-0.02	0.70	-0.02	0.76
Female	-0.01	0.82	-0.04	0.60
Children Under 16	-0.03	0.62	-0.05	0.50
Female & Child Under 16	0.03	0.64	-0.02	0.79
Female & Married	0.02	0.67	-0.01	0.93
Education	-0.00	0.96	-0.01	0.90
Temporary Contract	-0.04	0.48	0.01	0.88
Blue Collar	-0.04	0.44	-0.02	0.82
White Collar	0.15	0.01	0.02	0.82
Civil Servant	-0.28	0.00	0.00	1.00
German	0.04	0.52	0.02	0.84
West German	-0.22	0.00	-0.02	0.81
Satisfaction w Health	0.00	0.99	-0.08	0.27
Self-Reported Health Status	0.07	0.21	0.12	0.11

A.2 Balance Tables for Second Application

Table 4: Standardized Differences and p-values for Treated to Control Match in the Post-treatment Period for the Disability Payments Application

	Before Matching		After Matching	
	Std Dif	P-val	Std Dif	P-val
Regional Unemp.	0.12	0.02	0.01	0.88
Hourly Wage	-0.10	0.06	0.07	0.26
Age	0.01	0.86	0.11	0.11
Married	0.05	0.34	0.06	0.34
Female	0.03	0.55	0.04	0.55
Children Under 16	0.10	0.04	0.02	0.79
Female & Child Under 16	0.18	0.00	0.02	0.76
Female & Married	0.09	0.09	0.03	0.61
Education	-0.00	0.99	-0.04	0.51
Temporary Contract	0.10	0.08	0.10	0.11
Blue Collar	-0.02	0.77	-0.02	0.79
White Collar	0.12	0.02	0.02	0.79
Civil Servant	-0.24	0.00	0.00	1.00
German	-0.08	0.12	-0.02	0.73
West German	-0.17	0.00	-0.02	0.78
Satisfaction w Health	-0.02	0.62	-0.00	0.97
Self-Reported Health Status	0.02	0.64	0.05	0.41

Table 5: Standardized Differences and p-values for Pair-to-Pair Match in the Disability Payments Application

	Before Matching		After Matching	
	Std Dif	P-val	Std Dif	P-val
Regional Unemp.	0.06	0.40	-0.02	0.76
Hourly Wage	-0.28	0.00	-0.05	0.48
Age	-0.07	0.30	0.04	0.58
Married	-0.11	0.13	-0.03	0.71
Female	0.06	0.41	0.02	0.81
Children Under 16	-0.09	0.21	-0.04	0.65
Female & Child Under 16	-0.07	0.31	-0.05	0.52
Female & Married	-0.05	0.49	-0.04	0.64
Education	0.12	0.09	-0.02	0.82
Temporary Contract	0.13	0.06	0.05	0.54
Blue Collar	0.08	0.24	-0.02	0.84
White Collar	-0.08	0.27	0.02	0.84
Civil Servant	-0.02	0.81	0.00	1.00
German	0.06	0.36	0.06	0.47
West German	-0.05	0.48	0.01	0.89
Satisfaction w Health	0.21	0.00	0.05	0.47
Self-Reported Health Status	-0.04	0.53	0.04	0.60

Table 6: Standardized Differences and p-values for Treated to Control Match in the Pre-treatment Period for the Election Day Registration Application

	Before Matching		After Matching	
	Std Dif	P-val	Std Dif	P-val
Age	0.00	0.98	-0.05	0.19
African-American	-0.31	0.00	0.04	0.13
Female	-0.01	0.72	-0.04	0.28
Education	0.07	0.02	-0.05	0.20
Income	0.02	0.52	-0.05	0.18
Education X Income	0.07	0.02	0.05	0.19

Table 7: Standardized Differences and p-values for Treated to Control Match in the Pre-treatment Period for the Election Day Registration Application

	Before Matching		After Matching	
	Std Dif	P-val	Std Dif	P-val
Age	-0.06	0.04	0.05	0.15
African-American	-0.24	0.00	-0.04	0.13
Female	-0.01	0.77	0.04	0.25
Education	0.17	0.00	-0.05	0.23
Income	0.11	0.00	0.05	0.17
Education X Income	0.16	0.00	-0.05	0.19

Table 8: Standardized Differences and p-values for Treated to Control Match in the Pair-to-Pair Match for the Election Day Registration Application

	Before Matching		After Matching	
	Std Dif	P-val	Std Dif	P-val
Age	0.18	0.00	-0.05	0.29
African-American	-0.05	0.15	-0.07	0.12
Female	0.10	0.00	-0.02	0.67
Education	-0.27	0.00	0.05	0.28
Income	-1.27	0.00	-0.05	0.15
Education X Income	-1.10	0.00	-0.05	0.18