Stronger Instruments By Design*

Luke Keele† Jason Morgan‡

March 27, 2013

Abstract

Natural experiments provide one means for credibly estimating causal effects with observational data. The instrumental variable (IV) method is often applied to natural experiment reflecting a belief that combining the power of natural random assignment with an instrumental variable approach will solve many of the problems endemic to observational data. While IV analysis can be quite powerful, they also rest on a series of strong assumptions that may not be credible within a specific natural experiment. Here, we highlight how the bias from weak instruments is amplified when instruments are not as-if randomly assigned. We demonstrate how using an IV estimator based on matching and randomization inference can both correct for departures from as-if random assignment and strengthen the instrument. Specifically, we combine a matching algorithm with a reverse caliper and penalties to strengthen the instrument within a subset of the overall study population. We also demonstrate how researchers can probe the random assignment of the instrument assumption with a sensitivity analysis. We provide substantive examples of the proposed approach with a reevaluation and extension of a paper that uses rainfall as an instrument for voter turnout in U.S. counties (Hansford and Gomez 2010).

While randomized experiments have become prominent in political science, the discipline is replete with examples where experiments are infeasible. When experimentation is not possible, one alternative is to search for "natural experiments" where some intervention is thought to occur in an as-if random fashion, thus approximating a randomized experiment. Analysts search for such "gifts of nature" hoping to estimate unbiased effects (Rosenzweig and Wolpin 2000, pg. 872), and the estimates from natural experiments are often viewed as a close second best to a true randomized experiment (Dunning 2010).

*We thank Mike Baiocchi, Dylan Small and Paul Rosenbaum for helpful comments and discussion.

†Associate Professor, Department of Political Science, 211 Pond Lab, Penn State University, University Park, PA 16802, Phone: 814-863-1592, Email: ljk20@psu.edu

‡Ph.D. Candidate, Department of Political Science, 2140 Derby Hall, Ohio State University, Columbus, OH 43210 Email: morgan.746@osu.edu

There is much to be offered by natural experiments where some units obtain treatment and other are denied treatment in some haphazard manner. The difficulty, of course, is that haphazard assignment to treatment if often a far cry from a randomized experiment where randomization is a known fact. In the application we address below, we exploit rainfall patterns. While these patterns are certainly haphazard on any given day, they are also not randomly assigned in a way that the probability of both rain and sunny skies on election day is equal for all voters.

Due to the haphazard nature of treatment assignment, natural experiments often need support from statistical methods. The instrumental variables (IV) method is one statistical technique often used to exploit natural experiments. However, it is well known that the IV method has a number of strong assumptions that must hold for valid causal inferences (Angrist, Imbens and Rubin 1996). In this paper, we highlight how two of these assumptions can interact in a way that forms a particular threat to the use of IV with natural experiments. Specifically, we review how the problems caused by weak instruments are exacerbated when assignment of the instrument is not as-if random, which is common natural experiments (Small and Rosenbaum 2008). Next, we introduce a particular form of matching that allows us to pair units that are close on observables but far apart in terms of the encouragement induced by the instrument (Baiocchi et al. 2010). This procedure produces matches where assignment of the instrument should be ignorable with respect to observables but the effect of the instrument on the treatment should be as strong as possible, consequently strengthening the effect of the instrument. We apply this method to a study where rainfall is used as an instrument for turnout in the 2000 presidential election. We show that by selecting a subset of the units through matching, we can strengthen IV estimates from natural experiments and reduce sensitivity to bias from hidden confounders.

1 Building Stronger Instruments

We begin with a brief overview of instrumental variables as applied to natural experiments and then focus on two of the IV assumptions and how they can interact. We begin with a discussion of encouragement designs, which provide a template for many natural experiments.

1.1 IV and Its Assumptions

In the randomized encouragement design, some subjects are randomly encouraged to accept treatment, but some subset of the subjects fail to comply with the encouragement. IV provides an estimate of the treatment effect for those who take the treatment. The canonical example of the encouragement design is an experiment where some participants are encouraged to exercise. In this design, the instrument is encouragement to exercise, which is randomly assigned to study participants, the treatment is exercise, and the outcome of interest is lung function. Some subjects are randomly encouraged to exercise, but only some of those encouraged actually comply. Later all participants are measured on the outcome. The method of instrumental variables can be used to estimate the effect of the treatment, exercise, as opposed to the effect of encouragement (Angrist, Imbens and Rubin 1996). The IV estimate is often referred to as the complier average causal effect (CACE).

As applied to natural experiments, instrumental variable analysis is meant to mimic the randomized encouragement design (Holland 1988). For the IV method to provide valid causal inferences with natural experiments, five assumptions outlined by Angrist, Imbens and Rubin (1996) must hold.[1] These being: (1) random assignment of instrument status; (2) stable unit treatment value assumption; (3) no direct effect of instrument on outcome also known as the exclusion restriction; (4) monotonicity; and (5) no weak instruments. Below, we focus on assumptions 1 and 5 and focus on recent research which outlines how these two assumptions can interact particularly in the context of natural experiments.

Assumption (1) says that assignment of the instrument must be as-if random. For exam-

[1] See Sovey and Green (2011) for a more detailed introduction to IV analysis assumptions.

ple, in the encouragement design example, so long as the investigator assigns encouragement status through some random mechanism, such as a coin flip, this assumption will hold. In natural experiments, it is often unclear that the assumption of random assignment of the instrument holds since assignment to encouragement is not done in the controlled environment of a randomized experiment, but instead happens through some natural, haphazard process. It is for this reason that analysts should report whether encouraged and non-encouraged units are balanced with respect to observed covariates. A lack of balance would suggest that units with a particular covariate profile would be more likely to be assigned the instrument and serves as evidence that this assumption may have been violated (Dunning 2009; Sovey and Green 2011). For any natural experiment, the possibility always remains that the instrument is not randomly assigned. Analysts can use a sensitivity analysis to observe whether study conclusions are sensitive to this assumption (Rosenbaum 2010, 2002*b*, ch. 5), but unfortunately such sensitivity analyses are rare.

Next we focus on assumption (5) which states that analysts must avoid the use of a weak instrument. An instrument is said to be weak if manipulation of the instrument has little effect on treatment (Staiger and Stock 1997). In the encouragement design example, the instrument is weak if few of those encouraged to exercise actually exercise. The most common method of estimation used with instrumental variables is two-stage least squares (2SLS), and, as is well known, 2SLS can produce highly misleading inferences in the presence of weak instruments (Bound, Jaeger and Baker 1995). IV estimation with 2SLS takes identification of the IV estimand as given, and asymptotic approximations for standard errors and confidence intervals can incorrectly suggest strong treatment effects even when effects are nonexistent. Alternative methods of estimation are not prone to this problem. Rosenbaum (2005*b*) outlines an IV estimator based on permutation inference, where if the instrument is weak the confidence intervals will appropriately detect weak or non-existent identification. See Betz (2013) for one recent application of this IV estimator in political science.

Assumptions (1) and (5) are well known. What is less appreciated is that violations of these assumptions can interact to further hinder one's ability to make correct inferences with IV. Small and Rosenbaum (2008) prove important links between weak instruments and whether instrument status is as-if randomly assigned. If the instrument is randomly assigned but is weak, inferences from 2SLS can be misleading; though, with enough data, consistent estimates are possible. That is, with a weak instrument if Assummption (1) holds, the power to detect causal effects will be low. However, the key problem highlighted by Small and Rosenbaum (2008) is that even in a study with a very large sample size, a weak instrument is very sensitive to whether the instrument is as-if randomly assigned. In other words, when the instrument is weak even small departures from Assumption (1) can produce very large biases *no matter how large the sample size*. Small and Rosenbaum (2008) also prove that a strong instrument is more robust to departures from random assignment even in smaller sample sizes. Thus they show that if Assumption (1) does not hold, a smaller study with a stronger instrument is much less sensitive to bias than a weak instrument used in a much larger study.

This result has important consequences for applications of IV estimators to natural experiments. In a natural experiment, given that assignment to the instrument is haphazard instead of random, we must assume that deviations from Assumption (1) are common if not the rule. This implies that weak instruments pose an important threat to using IV in the analysis of natural experiments. Next, we present a design-based strategy for dealing with both issues.

1.2 Instrumental Variable Estimated via Matching

Our strategy relies on IV estimation via matching and randomization inference (Rosenbaum 1996, 2002*a*). Below, we begin with a formal description of IV estimation with matching and randomization inference. As we highlight below, IV with matching and randomization inference has important advantages over IV estimation via 2SLS.

The notation we introduce is for a paired randomized encouragement design. It is this experimental design that IV with matching mimics. There are I matched pairs, $i = 1, \ldots, I$, and the units within matched pairs are denoted with $j \in \{1, 2\}$. We form these pairs based on observed covariates, $\mathbf{x}_{ij}$, which are measured before assignment to the instrument. After matching, we assume $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for $i = 1, \ldots, I$. One subject in each matched pair is encouraged to accept a high dose of treatment, which we denote as $Z_{ij} = 1$, and the other subject is not encouraged denoted by $Z_{ij} = 0$, so that $Z_{i1} + Z_{i2} = 1$ for $i = 1, \ldots, I$. If $Z_{ij} = 1$ the unit is encouraged and subject ij receives treatment at dose d_{Tij} and if $Z_{ij} = 0$ subject ij receives treatment at dose d_{Cij}. Consistent with the potential outcomes framework (Neyman 1923; Rubin 1974), these doses, however, are potential quantities which implies that we do not observe the pair (d_{Tij},d_{Cij}). We do observe the dose actually received, which is $D_{ij} = Z_{ij} d_{Tij} + (1 - Z_{ij}) d_{Cij}$. Each subject also has two potential responses which we denote as r_{Tij} if $Z_{ij} = 1$ or r_{Cij} if $Z_{ij} = 0$. As with the doses, we do not observe the pair of potential outcomes: (r_{Tij}, r_{Cij}), but we do observe the responses: $R_{ij} = Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij}$.

Following Rosenbaum (1996), we assume that the effect of encouragement on response is proportional to its effect on the treatment dose received,

$$r_{Tij} - r_{Cij} = \beta(d_{Tij} - d_{Cij}). \tag{1}$$

If this model is true then observed response is related to observed dose through the following equation

$$R_{ij} - \beta D_{ij} = r_{Tij} - \beta d_{Tij} = r_{Cij} - \beta d_{Cij}. \tag{2}$$

Under this model of effects, the response will take the same value regardless of whether or not $Z_{ij} = 1$ or $Z_{ij} = 0$ in each pair. This implies that this model of effects satisfies the exclusion restriction. Informally, the exclusion restriction implies that instrument assignment Z_{ij} is related to the observed response $R_{ij} = Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij}$ only through realized dose of the treatment D_{ij}. That is true here since $R_{ij} - \beta D_{ij}$ is a constant that does not vary with Z_{ij}.

Given this model of effects, we wish to test whether the treatment is without effect, estimate a point estimate, and form a confidence interval for this point estimate. Under randomization inference, we can calculate these quantities by testing various hypotheses about β using the following set of null hypotheses $H_0 : \beta = \beta_0$. We obtain exact inferences about β using the observed quantity $R_{ij} - \beta_0 D_{ij} = W_i$ as a set of adjusted responses.

To test the sharp null hypothesis, we test $H_0 : \beta = \beta_0$, with $\beta_0 = 0$ by ranking $|W_i|$ from 1 to I. We calculate Wilcoxon's signed rank statistic, W_{β_0}, as the sum of the ranks for which $W_i > 0$. If ties occur, average ranks are used as if the ranks had differed by a small amount. Under $H_0 : \beta = \beta_0$, if $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for each i, and there are no unobserved confounders related to the probability of treatment, then the probability of assignment to the instrument is 1/2 independently within each pair. If this is true, we can compare W_{β_0} to the randomization distribution for Wilcoxon's signed rank statistic and this provides the exact p-value for a test of the sharp null hypothesis for $\beta_0 = 0$.

A point estimate for β is obtained using the method of Hodges and Lehmann (1963). The estimate of β is the value of β_0 such that Wilcoxon's signed rank statistic is as close as possible to its null expectation, $I(I+1)/4$. Intuitively, the point estimate $\hat{\beta}$ is the value of β_0 such that W_{β_0} equals $I(I+1)/4$ when W_{β_0} is computed from $R_{ij} - \beta_0 D_{ij}$. It is not necessary to use the signed-rank statistic. Rosenbaum (2005*b*) shows that if we use the sample mean of W_i this yields

$$\hat{\beta} = \frac{\sum_{i=1}^{I}(2Z_{i1} - 1)(R_{i1} - R_{i2})}{\sum_{i=1}^{I}(2Z_{i1} - 1)(D_{i1} - D_{i2})},$$

which is the well-known Wald (1940) estimator.

A 95% confidence interval for the treatment effect is formed by testing a series of hypotheses $H_0 : \beta = \beta_0$ and retaining the set of values of β_0 not rejected at the 5% level. This is equivalent to inverting the test for β (Rosenbaum 2002*b*).

One advantage of applying randomization inference to IV becomes apparent with interval estimation when instruments are weak. An instrument is strong if d_{Tij} is considerably larger than d_{Cij} for most or all individuals ij. An instrument is weak if d_{Tij} is close to or equal to d_{Cij} for most individuals ij. In other words, an instrument is weak instrument when most units ignore the encouragement to take the treatment. When the instrument is weak, two-stage least squares often provides incorrect inferences where 95% confidence intervals may over cover 85% of the time (Bound, Jaeger and Baker 1995). The confidence intervals from randomization inference have exact 95% coverage regardless of whether the instrument is strong or weak (Imbens and Rosenbaum 2005). Under randomization inference, if the instrument is weak, the interval becomes longer and perhaps even infinite in length. In this case, a long confidence interval is a warning that the instrument provides little information about the treatment. In the context of a natural experiment, a long confidence interval also implies that special attention should be paid to whether the instrument is balanced.

1.3 Sensitivity Analysis

The nonparametric IV estimator is based on the assumption that assignment to encouragement, $Z_{ij} = 1$ or $Z_{ij} = 0$, is constant within pairs as if encouragement status were determined by flipping a coin. When we suspect assignment of the instrument is not fully as-if random, a sensitivity analysis can be performed to assess how departures from random assignment of the instrument might alter our conclusions (Rosenbaum 2002b).

For two subjects, i and j matched so that observed covariates are identical, $\mathbf{x}_{ik} = \mathbf{x}_{ij}$, we assume that probability of treatment is constant within matched pairs. However, subjects may differ in the probability of treatment because they differ in terms of some unobserved covariate. Let π_j denote the probability of treatment for unit j. We may characterize this probability with a logit model linking the probability of treatment to observed covariates $\mathbf{x}_j$ and an unobserved binary covariate u_j: $\log\{\pi_j/(1-\pi_j)\} = \phi(\mathbf{x}_j) + \gamma u_j$ where $\phi(\cdot)$ is an unknown function. Using this model, we can express how two matched units might differ in terms of their probability of treatment as a function of u. For two units, ik and ij with $\mathbf{x}_{ik} = \mathbf{x}_{ij}$, we characterize how they may differ in their odds of treatment with the model above rewritten as: $\pi_{ij}(1-\pi_{ik})/\pi_{ik}(1-\pi_{ij}) = \exp\{\gamma(u_{ij} - u_{ik})\}$ (Rosenbaum 2002b, sec.

4.2). Now we write $\exp(\gamma) = \Gamma$, and if $\Gamma = 1$ for two matched units, then the units do not differ in their odds of treatment as a function of the unobserved u.

We can vary the values of Γ systematically as a sensitivity parameter to probe whether the IV estimate is sensitive to departures from random assignment of the instrument. For Γ values greater than one, we can place bounds on quantities of interest from the IV estimator. For example, consider the p-value for $\hat{\beta}$. First, we assume $\Gamma = 1$ and apply randomization inference. The resulting p-value is valid if there are no hidden confounders. In the sensitivity analysis, we pick a range of Γ values and see at what value of Γ our p-value exceeds 0.05. For example, if we find that the estimated p-value exceeds 0.05 when $\Gamma = 1.05$ this suggest that a very slight departure from randomization of the instrument might overturn our study. If, on the other hand, we find that the p-value from our study exceeds 0.05 when Γ is greater than say 4, this suggests that unless the departure from randomization of the instrument is fairly substantial our conclusions would still hold. That is, the odds of treatment would have to differ by a factor of 4 before our conclusions would be overturned. A sensitivity analysis is especially important in the context of a natural experiment where assignment of the instrument is haphazard and may not be as-if random. In the analysis, here, Γ becomes an important outcome measure. Strengthening the instrument should make the IV estimate less sensitive to bias from hidden confounders. As such, in the analyses that follow we compare Γ values to see whether stronger instruments are associated with higher Γ values.

1.4 Strengthening Instruments Through Matching

Next, we demonstrate how we can alter the matching process to also create stronger instruments. Thus far we have assumed that the encouragement, Z_{ij}, is binary. When the instrument is binary, subjects are either encouraged or not, and there is no room to strengthen the encouragement provided by the instrument. However, many instruments are continuous. We now assume that the instrument is a continuous measure where higher values represent a greater level of encouragement; i.e., a higher dose of the treatment is more likely for higher values of Z_{ij}. With a continuous instrument, we might imagine that the ideal matched pair of subjects ik and il would have the following property $\mathbf{x}_{ik} = \mathbf{x}_{il}$ but $Z_{ik} - Z_{il}$ should be large. That is, these units should be identical in terms of observed covariates but one of the units is strongly encouraged to take a high dose of the treatment while the other is not. Such a match ensures comparable units with a large difference in terms of encouragement allowing for a strong instrument. How might we implement such a match?

This ideal match creates complications for typical matching algorithms that are applied to data with identifiable treatment and control groups define by a binary indicator. With a continuous instrument, clearly identifiable groups for the matches don't exist. The most obvious way to apply a standard matching method would be to divide the observations into two groups based on some cutoff in the instrument. However, dividing the data in this way requires an arbitrary choice about where to split units into the encouraged and unencouraged groups. We can solve this problem with optimal nonbipartite matching (Greevy et al. 2004; Lu et al. 2011). Nonbipartite matching can be used when the data are not explicitly divided into treatment and control groups. Instead, all of the observations are pooled and matched pairs are formed so that the sum of pairwise distances is minimized.

Baiocchi et al. (2010) demonstrate how to use nonbipartite matching to implement the ideal IV match. We motivate their method through the concept of a reverse caliper. Caliper matching is a method that attempts to avoid poor matches by imposing a tolerance on the maximum distance between matched pairs (Cochran and Rubin 1973). For two subjects i and j, let P_i and P_j be a score on a distance metric such as the propensity score. Under a caliper, a match for subject i is selected only if $||P_i - P_j|| < \Lambda$, where Λ is a pre-specified tolerance. We can reverse the concept of a caliper and say a match for subject i is selected only if $||P_i - P_j|| > \Lambda$, where Λ remains a pre-specified tolerance. How might this help us build stronger instruments? We use the reverse caliper in conjunction with matching to find units that are similar on observables, but we only keep a matched pair if it satisfies the reverse caliper such that the matched pair is very dissimilar in terms of the encouragement provided by the instrument.

To enforce the reverse caliper, we use penalties. We added a substantial penalty to the discrepancy between any pair of units whose distance on the instrument differed by less than Λ. Penalties are used to enforce compliance with a constraint whenever compliance is possible, and also to minimize the extent of deviation from a constraint whenever strict compliance is not possible. Thus the matching algorithm attempts to balance observables subject to the constraint on instrument distance. More formally, let z_{kl} denote the value of the instrument for unit l in possible pair k, and $z_{k1} - z_{k2}$, the distance between the observations in the matched pair on the instrument. A distance penalty, p_k, is defined as

$$p_k = \begin{cases} (z_{k1} - z_{k2})^2 \times c & \text{if } z_{k1} - z_{k2} < \Lambda \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where Λ is some threshold pre-defined by the analyst. The penalty, p_k, is defined such that a smaller value of $z_{k1} - z_{k2}$ receives a larger penalty making those two units less likely to be matched, while c scales the penalty to that of the distance matrix. See Rosenbaum (2010, Sec. 8.4) for a discussion of penalties in matching.

To be fully effective, however, the penalized matching must also be combined with what are known as "sinks" (Lu et al. 2001). Matching with penalties alone tends to produce some matched pairs that are distant on the instrument but have suboptimal balance on observables. To avoid such matches, we discard suboptimal pairs with sinks. To eliminate e units that create the suboptimal matches, e sinks are added to the data before matching. We define each sink so that it has a zero distance between each unit and an infinite distance to all other sinks. This will create a distance matrix of size $(2I + e) \times (2I + e)$. The optimal nonbipartite matching algorithm pairs e units to the e sinks in such a way to minimize the total distance between the remaining $I - e/2$ pairs. That is by pairing a unit with a sink, the algorithm removes the e units that would form the e set of poor matches. Thus the best possible set of e units are removed from the matches.

The matching algorithm, then, creates the optimal set of matched pairs that are similar in terms of covariates but differ in levels of encouragement. Moreover, any matched units for which it's difficult to balance and increase distance on the instrument are excluded from the study. This leads to a smaller study in hopes of strengthening the plausibility of the IV analysis. This form of matching is consistent with efforts to focus on smaller homogeneous subsets of the data because comparability is improved and sensitivity to unobserved bias is lessened (Rosenbaum 2005*a*).

1.5 The Estimand

Anytime one analyzes a natural experiment, it is worth carefully considering the estimand: the quantity being estimated. Often the estimand in a natural experiment may not be the quantity of interest in a theory of politics (Sekhon and Titiunik 2012; Keele and Minozzi 2012). This is especially true when IV is applied to a natural experiment, since IV only identifies what is typically know as the complier average causal effect (CACE) which is the average causal effect among those who comply with the encouragement to the treatment (Angrist, Imbens and Rubin 1996). Often this estimand is referred to as the local average treatment effect (LATE). The strategy we use here further changes the estimand by increasing the locality of the IV estimand. That is, we restrict the IV estimator to a subpopulation that not only responds to the encouragement but for whom encouragement is strong. In short, our estimand is more local than the usual LATE. We argue that it is better to restrict the analysis to units for whom we have greater confidence that identification holds. While we may not be able to easily extrapolate our estimates to other populations, we argue that interval validity comes first in the process of causal inference. There is no reason to worry about extrapolation of effects when we have little confidence in the identification strategy.

2 Application: Turnout and Democratic Vote Share in the 2000 US Presidential Election

We demonstrate these techniques through an analysis of a long-running debate in the American politics literature: whether turnout affects election results. The question of interest is whether higher levels of turnout would increase vote share for the Democratic party. Of course, any estimates of the effect of turnout on vote share are subject to endogeneity. Hansford and Gomez (2010) exploit a natural experiment created by rainfall patterns on election day to overcome this problem. Based on evidence that rainfall depresses turnout (Gomez, Hansford and Krause 2007), they use rainfall as an instrument for turnout. They find that higher turnout does help Democratic candidates.[2]

Before, we turn to our analysis, we evaluate the plausibility of the IV assumptions in this context focusing specifically on Assumptions (1) and (5). The design does fit the encouragement design paradigm, as rainfall discourages the treatment of voting. Moreover, the exclusion restriction seems plausible. While the exclusion restriction is important, we focus on Assumptions (1) and (5), which are often overlooked by analysts. For example, Hansford and Gomez (2010) do not discuss whether assignment of the instrument is as-if random, and they provide only a brief discussion of whether the instrument is weak or not.

We begin with a discussion of the as-if random status of rainfall. In their analysis, Hansford and Gomez (2010) use deviations from average rainfall on election day as their instrument. So the design studies whether unusually rainy weather discourages turnout. This design decision bolsters Assumption (1) since relative amounts of rainfall is more likely to be random than simple rainfall patterns, which are correlated with region and perhaps political characteristics. As in many natural experiments, however, the validity of Assumption (1) is hard to assess *a priori*, and we should rely on empirical evidence to judge the validity of

[2]Hansford and Gomez assess four different hypotheses related to turnout. For the purposes of this analysis, we focus on whether increased turnout increases Democratic presidential candidate vote share. The other three hypotheses tested by Hansford and Gomez include whether the size of this effect depends on the partisan composition of the electorate, if turnout decreases incumbency vote share, and whether turnout increases electoral volatility.

this assumption. If rainfall is a valid instrument, we should expect that characteristics like levels of education and income should be balanced across counties with unusual amounts of rain on election day and those counties with normal precipitation patterns on election day. As we demonstrate below, this is not the case, and we must adjust for covariate imbalances. Moreover, since adjustments are required, we also conduct a sensitivity analysis to understand whether small departures from random assignment of the instrument might affect our conclusions.

Next, we ask whether the instrument is weak? Hansford and Gomez (2010) note that rainfall seem unproblematic based on a weak instrument test where the F-statistic passes the usual rule-of-thumb threshold of 10 (Sovey and Green 2011). However, careful evaluation of the design reveals a flaw that suggests the rainfall instrument is actually weak. In the design, unusual amounts of rain are thought to discourage turnout among Democrats. The instrument is measured as the difference between rainfall on election data and the average rainfall amounts for the county during the week of the election. This measure of rainfall deviations is, of course, two-sided. Some places experience more rain than is typical on election day, while other places have better than average weather. The power of the instrument in this design, however, stems from the discouragement provided by abnormally large amounts of rain, not from drier than average weather. In the analysis, however, the measure used is the raw measure of both positive and negative deviations. Why is this a concern? In the original diagnosis on the problems caused by weak instruments, Bound, Jaeger and Baker (1995) found that using randomly generated noise with no information about the treatment as an instrument provided estimates that appeared to be identified. Thus one lesson from that study is that additional variation in the instrument, even stochastic variation unrelated to the treatment, may provide false identification. We contend that the extra variation in the instrument created by the better than average weather may make this instrument appear stronger than it really is.

To that end, we performed a reanalysis to understand whether using the full range of rainfall deviations may affect the weak instrument test. First, we replicated the weak instrument test results from the original analysis. We found the F-statistic from the weak instruments test was 20.6 well above the rule-of-thumb threshold of 10. In a new analysis, we altered the rainfall instrument such that counties with better than average weather on election day were coded as zero, since we expect such weather patterns to provide little in the way of discouragement to vote. We then performed a new weak instrument test using the new version of the instrument. Now the F-statistic from the weak instruments test is 2.4 well below the usual threshold of 10. Our reanalysis suggests that the rainfall instrument is relatively weak and the original test results may have been a function of the extra variation provided by better than average weather. In short, for this application, rainfall patterns are no as-if random and the instrument is weak. Next, we apply matching to increase balance and strengthen the instrument.

2.1 County Level Analysis

We do not perform a complete reanalysis of Hansford and Gomez (2010). Instead, we restrict our analysis to the 2000 presidential election. We chose the 2000 presidential election for three reasons. First, we use a more recent election since it allows us to utilize a wider number of covariates when we evaluate whether the rainfall instrument created as-if random status. Second, we use 2000 presidential election since there was a large variation in rainfall on election day that year. Finally, we think it is useful to restrict the matched county pairs to a single year to reduce heterogeneity and bias from unobserved confounders. As in the original paper, we also exclude Southern counties from the data to increase the comparability between the original analysis and our analysis.[3]

We use the county-level data provided by Hansford and Gomez (2010) in their replication data set. These data include election day rainfall measured relative to the average rainfall around election day; presidential election turnout, measured as a percentage of votes cast for

[3]For the reasoning behind this, see their detailed discussion of the data (Hansford and Gomez 2010, pp. 273-276).

presidential candidates divided by the voting age population; and the two-party vote share for the Democratic presidential candidate. Overall, the data set includes more than 1900 counties across 36 US states.

In our analysis, we added several covariates that are likely to be related to turnout and electoral outcomes. These covariates include the natural log of county population, the proportion of black and Hispanic residents in the county, educational attainment measured by the percentage of high school and college educated residents, the natural log of median household income, and the county poverty rate. It is these covariates that should be balanced if the rainfall instrument is as-if random.[4] We might also consider whether turnout and vote share in 1996 or 1998 were also balanced, but since these measures could be affected by the instrument (rainfall in previous years) and treatment (turnout in previous elections), we exclude measures of this type from the analysis.

Before the analysis, we first tested whether the assignment of rain on election day in 2000 produced as-if random conditions. We do that by determining whether the covariates were balanced by rainfall patterns. All counties that experienced greater than normal rainfall were considered part of the treated group and all other counties were considered to be the control. There were 1233 counties (64% of the sample) in the treated group and 692 counties (36% of the sample) in the control group. We then conducted balanced tests on the Census data. In Table 1 for each covariate we report means, the standardized difference in means, and the p-value from the Kolmogorov-Smirov (KS) test.[5] We find that in fact, rainfall patterns in 2000 were not as-if random. While the means of some covariates are quite close, we see that treated counties had lower levels of college education and income. For example, the percentage of college educated residents was four percent lower in treated counties than in control counties. As is often the case, however, simply examining means hides differences in other moments of the distributions. We find in every case that the p-values from the KS test are below 0.001. Coupled with the fact that rainfall is a relatively weak instrument for turnout—the

[4] Summary statistics for all of the data used in this analysis are presented Appendix ??.
[5] Balance statistics reported calculated in R with the Matching package (Sekhon 2011).

Table 1: Balance statistics for unmatched US covariates

	Mean treated	Mean control	Std. diff.	KS p-val[a]
Rainfall deviation	0.34	−0.07	1.38	0.00
Population (log)	9.99	10.56	0.38	0.00
Percent African-American	0.02	0.03	0.18	0.00
Percent Hispanic	0.03	0.07	0.42	0.00
Percentage with High School Degree	0.80	0.82	0.24	0.00
Percentage with College Degree	0.16	0.20	0.59	0.00
Median Household Income (log)	10.45	10.52	0.30	0.00
Percentage Below Poverty Line	0.13	0.13	0.00	0.00

[a] Kolmogorov-Smirnov p-values calculated by with $b = 5000$ bootstrap replications.

correlation between rainfall and turnout is 0.08 in 2000—it would be dangerous to make inferences based on the standard 2SLS estimator. We now turn to a matching analysis that attempts to remedy these issues.

In our analysis, we applied optimal nonbipartite matching to create two sets of matches. In the first match, we simply attempted to balance observables, but we did not enforce any minimum discrepancy on the matched pairs in terms of rainfall. In the second match, the objective is to maximize balance, while also maximizing distance on the instrument. For both matches, we calculated the pairwise distances between the counties included in the sample. For our application we used a rank-based Mahalanobis distance metric, which is robust to highly skewed variables. This produced a symmetric 1925×1925 distance matrix. To control for possible SUTVA violations across contiguous counties, we then applied a large penalty for geographic proximity. With a large penalty for proximity, the algorithm will avoid matches for contiguous counties if possible subject to minimizing imbalances on the covariates.[6]

For the second match, we applied the reverse caliper through a penalty function to all potential matched pairs where the distance on the instrument was less than a given threshold. As we expected, maximizing distance on the instrument tended to worsen balance. To that

[6] The penalty was large enough that it basically precluded the possibility of matching contiguous counties.

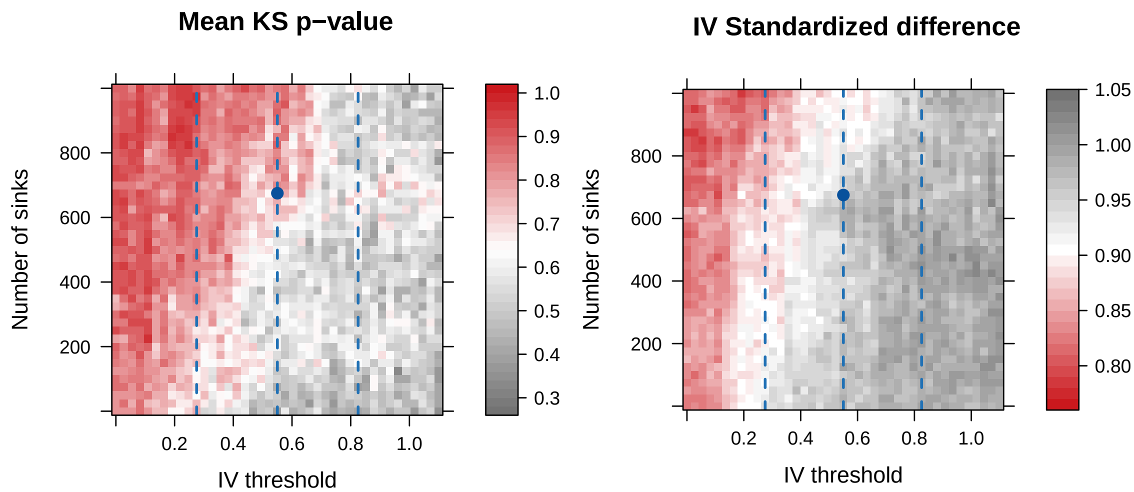

Figure 1: Balance on covariates and instrument: 1681 diagnostic matches for US county data. Dashed vertical lines indicate 1, 2, and 3 standard deviations for the pairwise distance on the instrument. Shading indicates the level of balance and the dot shows the parameter values selected for further analysis.

end, we simultaneously introduced a number of sinks to counteract the negative impact the instrument distance penalty had on overall covariate balance. To determine and optimal threshold for the value of the reverse caliper and the number of sinks, we used a grid search for values of the threshold from 0 to 1.10 (4 times the standard deviation of the pairwise distances on the instrument) and sinks between 0 and 1000 (dropping approximately 50% of the observations). Overall, 41 different values of the threshold and sinks were used, producing 1681 different match specifications.[7] For each specification, optimal nonbipartite matching was performed and balance statistics on the covariates and instrument were recorded. These balance results are summarized graphically in Figure 2.1.

Figure 2.1 illustrates the trade-off between overall covariate balance and the strength of the instrument.[8] When we don't place a constraint on the strength of the instrument, balance

[7]There is nothing special about the use of 41 different values of the threshold and sinks. It was simply a convenient number that produced interpretable values on the threshold.

[8]Here we discuss the mean p-value for Kolmogorov-Smirnov two-sample tests on the seven covariates discussed above. The patterns we highlight hold for median QQ statistics, $t-$tests, and Wilcoxon signed rank statistics.

on the covariates is easy to achieve, but as the threshold is increased—as the penalty for being close on the instrument becomes more severe—overall balance declines. At the same time, increasing the threshold does have the desired effect of increasing the distance on the instrument in matched pairs. It is also clear from the figure that introducing sinks has the desired effect of improving balance. As the number of sinks included in the match increases, overall balance increases.

In this analysis, we choose a sinks/instrument threshold combination that prioritized covariate balance, while providing a modest increase in the strength of the instrument. Specifically, we chose a threshold equal to two standard deviations on the pairwise differences in rainfall, or $\Lambda = 0.55$, and 675 sinks which drops approximately one third of the observations. This match is indicated in Figure 2.1 as a dot.

Table 2 presents balance statistics from the two matches. The first set of balance statistics presented in the table, which we label Weak IV, is for the match that does not attempt to maximize the discrepancy between matched counties on rainfall deviation and does not utilize any sinks.[9] The second match, which we label Strong IV, includes both the reverse caliper and the sinks. That is, we strengthen the instrument by increasing distance on the instrument, but we also seek to maximize balance through the use of sinks. Both matches are well balanced, and one would not prefer one match over the other on the basis of this criterion. The Strong IV match, however, produces a match where the standardized difference on the rainfall measure increases from 0.82 to nearly 1.0. As such, we were only able to strengthen the instrument a modest amount. We found that stronger instruments produced levels of balance we thought unacceptable.

Of course, to increase the strength of the instrument we had to discard a number of observations. The Strong IV match uses 625 matched county pairs instead of full set of 962 matched county pairs used in the Weak IV match. Given that weak instruments create biases that outweigh sample size considerations, the Strong IV match is to be preferred (Small and

[9]The Weak IV match does include the penalty preventing the matching of geographically contiguous counties.

Rosenbaum 2008). We now turn to estimating the effect of interest and a sensitivity analysis. In both cases, we should be able to understand whether strengthening the instrument alters our inference.

Table 2: Balance statistics for US county matches

	(I) Weak IV[a] $I = 962$ matched pairs					(II) Strong IV[a] $I = 625$ matched pairs				
	Mean treated	Mean control	Std. diff.	Med. QQ	KS p-val[b]	Mean treated	Mean control	Std. diff.	Med. QQ	KS p-val[b]
Rainfall deviation	0.32	0.07	0.82	0.27	0.00	0.35	0.06	0.96	0.33	0.00
Population (log)	10.18	10.22	0.03	0.01	0.40	10.04	10.07	0.02	0.02	0.70
Percent African-American	0.02	0.02	0.00	0.01	0.97	0.02	0.02	0.01	0.01	0.58
Percent Hispanic	0.05	0.05	0.03	0.01	0.76	0.05	0.04	0.01	0.01	0.68
Percentage with High School Degree	0.81	0.81	0.01	0.01	1.00	0.81	0.81	0.03	0.01	0.83
Percentage with College Degree	0.17	0.17	0.04	0.01	0.83	0.17	0.17	0.00	0.00	0.99
Median Household Income (log)	10.48	10.48	0.01	0.01	0.86	10.46	10.46	0.02	0.01	0.78
Percentage Below Poverty Line	0.13	0.13	0.02	0.01	0.76	0.13	0.13	0.04	0.01	0.64

[a] Match (I) performed without reverse caliper or sinks; (II) performed with $\varepsilon = 0.55$ and 675 sinks.
[b] Kolmogorov-Smirnov p-values calculated from 5000 bootstrapped samples.

Table 3 contains the IV point estimate, a 95% confidence interval, and results from the sensitivity analysis for the two matches. We first focus on the results from the Weak IV match. For the Weak IV match, the point estimate is 0.40, which implies that an increase in turnout of one percentage point increases Democratic vote share by four-tenths of a percent. The point estimate is statistically significant as the 95% confidence interval is bounded away from zero. Next, we turn to the point estimate from Strong IV match. For the matches which resulted in a stronger instrument, we find that the point estimate is now 0.60 or 50% larger than from the design with the weaker instrument. Thus in counties where we are able to strengthen the instrument we find that if turnout goes up by one percentage point, then the Democratic vote share increase six-tenths of a percent. These estimates suggest that the modest gains in instrument strength produces a larger point estimate.

These estimates assume that assignment to an above average amount of rain on election day within matched pairs is effectively random. We next ask whether these estimates are sensitive to bias from a hidden confounder that alters the probability of being assigned to above average rain within matched pairs. We also focus on whether the design with the smaller sample size but stronger instrument is more resistant to bias from an unobserved confounder than the design with the weaker instrument. We find that the weaker instrument is more sensitive to biases from unmeasured confounders. With the weaker instrument, we find that $\Gamma = 1.17$ as compared to $\Gamma = 1.31$ for the design with the stronger instrument. For the Weak IV match, the sensitivity analysis implies that if a hidden confounder changed the odds of receiving an above average amount of rain on election day with matched pairs by a factor of .17 that would explain the observed association. For the Strong IV match, the odds of differential instrument assignment need to be 13% larger before we can explain the observed association. Thus the estimate from the Strong IV match is 13% less sensitive to unmeasured bias even though we used 337 fewer matched pairs than in the design with the weaker instrument.

Before moving to a second analysis, we turn to the interpretation of IV estimates for this application. Hansford and Gomez (2010) do not discuss how the use of IV changes the interpretation of their estimates, and we think it is worth carefully understanding what is being estimated. One might presume that the IV estimate is simply the effect of turnout on vote share purged of any endogenous variation. However, as we outlined earlier, IV provides an estimate of the LATE. Therefore, IV identifies the average causal effect of the treatment for the subset of the population whose selection into the treatment was induced by the instrument. Here, the IV estimand is the average causal effect for the subset of all counties that voted at a lower rate when subjected to an unusual amount of rainfall on election day but would have voted at a higher rate if it had not rained. Importantly, IV provides us with an average effect *only* among those induced to take the lower turnout treatment by the rainfall intervention. As such, the IV estimate only refers to those who respond to the rainfall instrument: the counties that are sensitive to unusual rainfall patterns on election day. In general, we might wonder about the political relevance of the estimand in this application. That is, we might ask whether we have any theory about counties that are sensitive to rainfall? We can identify which states contribute the largest percentage of counties influenced by rainfall. We found that states such as North Dakota, Minnesota, Iowa, Missouri, Kentucky, Illinois, Colorado, and Wisconsin contribute more than half of the treated counties. As we might suspect arid states like New Mexico, Arizona, California, and Nevada contribute few or no counties to the analysis. An analysis by Fraga and Hersh (2010) suggest these counties tend to be ones with uncompetitive elections. Moreover, the Strong IV design changes the estimated further. For this design, the effect is a more local version of LATE; one that only applies to counties that were strongly discouraged to vote by rain on election day. Full consideration of this question is beyond the scope of this article. However, when analysts apply IV to a natural experiment, they must seriously consider the interpretation of the LATE.

Table 3: Treatment effect estimates and sensitivity analysis for US counties match

	Weak IV			Strong IV		
	$\hat{\beta}$	95% CI		$\hat{\beta}$	95% CI	
	0.400	0.200	0.577	0.600	0.358	0.872

Γ	Weak IV	Strong IV
1.00	0.000	0.000
1.15	0.024	0.001
1.16	0.032	0.001
1.17	0.041	0.002
1.18	0.052	0.003
1.19	0.065	0.004
1.20	0.080	0.005
1.30	0.368	0.040
1.31	0.407	0.047
1.32	0.446	0.056
1.33	0.486	0.065

2.2 Precinct Level Analysis

We, next, turn to a second empirical example where we again use rainfall as an instrument for turnout. Now, however, we confine the analysis to a single Southern state, Georgia in 2000. This second example has three advantages over the prior one. First, as a Southern state, Georgia wasn't included in the original analysis, which excluded all Southern states. A study of Georgia, then, represents an out-of-sample test. Second, by looking at a single state we are able to conduct the analysis at a much finer level of geographic resolution. In this case, we extend the analysis to the precinct level. Finally, looking at a single state, utilizes a design that allows us to control for state-level confounders. Keele and Minozzi (2012) demonstrate that turnout is subject to state level confounders that are difficult to control for in an analysis with a national sample of counties. We selected Georgia since in 2000 there was considerable variation in rainfall on election day with some parts of the state receiving little rain while other parts of the state were unusually rainy.

The first step in extending the analysis to Georgia was gathering the requisite data on rainfall. We used the same EarthInfo weather station data used in the original analysis. To extend the rainfall measurements from weather stations to precincts, we used the same method as the original analysis: kriging. Kriging is a statistical technique that produces an estimate of the underlying (usually assumed to be smooth) surface by a weighted average of the data, with weights declining with distance between the point at which the surface is being estimated and the locations of the data points. In short, we estimate a rainfall surface based on kriging and use the predictions from that model to estimate rainfall on election day for each precinct.[10] We estimate rainfall for the centroid of each precinct. We did this for both rainfall on election day in 2000 and the average amount of rainfall in November. The analysis is based on readings from all the weather stations in Georgia as well as a stations from areas in bordering states that are close to the Georgia border. We included weather stations from nearby states since the kriging model is strengthened by incorporating data from near the Georgia border. For more rural parts of GA, estimating rainfall at the precinct level adds little information since most counties have only a single weather station. In the Atlanta metropolitan area and in other urban areas, however, the counties have multiple weather stations. Since the bulk of the precincts are in urban areas, we are able to add a much more fine grained estimates of local rainfall patterns on election data.

In this analysis, we match on the same covariates that we used in the county level analysis, with one exception. We did not include the percentage of Hispanics in each precinct, since there were very few Hispanics in Georgia in 2000. Census data in 2000 was collected at either the block or blockgroup level. We hired a GIS analyst to either aggregate the block level data to the precinct geography or estimate precinct level measures from the blockgroups since in Georgia precincts typically differ by less than 1% in terms of population. Thus all census measures were used as precinct level covariates.

Again, we start by examining whether the rainfall deviations measure appears to be

[10]For the kriging model, we used the K-Bessel function to model decay from each weather station. We estimate the model with and without anisotropy to allow for possible directionality in the estimates due to wind. The model fit with and without anisotropy was nearly the same. We checked whether our kriging model was invariant to a number of different model parameters, but found little change in the estimates.

allocated in an as-if random fashion. Table 4 contains balance statistics for the unmatched precincts. As before, allocation of the treatment does not meet the as-if random criterion. The treated counties had a lower percentage of African-Americans, higher education levels, lower poverty rates, and higher incomes. Testing the difference in terms of the distributions, we find that all the KS test p-values are below any reasonable threshold.

Table 4: Balance statistics for unmatched Georgia covariates

	Mean treated	Mean control	Std. diff.	KS p-value[a]
Rainfall deviation	1.35	−1.52	1.65	0.00
Population (log)	7.93	7.64	0.35	0.00
Percent African-American	0.24	0.30	0.20	0.00
Percentage with High School Degree	0.18	0.20	0.39	0.00
Percentage with College Degree	0.11	0.08	0.43	0.00
Median Household Income (log)	10.72	10.53	0.44	0.00
Percentage Below Poverty Line	0.11	0.15	0.33	0.00

[a] Kolmogorov-Smirnov p-values calculated by with $b = 5000$ bootstrap replications.

We again applied optimal nonbipartite matching to the precinct level data in Georgia. As before, we used the rank-based Mahalanobis distance metric and applied penalties to avoid SUTVA violations and repeated the iterative search for an optimal distance for the reverse caliper and the number of sinks. We used the same grid search for values of the threshold from 0 to 1.10 (4 times the standard deviation of the pairwise distances on the instrument) and sinks between 0 and 1500 (which corresponds to using all of the observations to dropping approximately 45% of the observations). For each specification, optimal nonbipartite matching was performed and balance statistics on the covariates and instrument were recorded. With the Georgia data, we are able to produce matches with a stronger instrument that preserves balance than we were in the county level analysis. With the Georgia data, we chose a threshold equal to two standard deviations on the pairwise differences in rainfall, or $\Lambda = 1.55$, and 1475 sinks. The Strong IV match uses about 45% of the precincts as compared to the Weak IV match, which uses every precinct.

Table 5: Balance statistics for Georgia precinct matches

	(I) Weak IV[a] $I = 1335$ matched pairs					(II) Strong IV[a] $I = 597$ matched pairs				
	Mean treated	Mean control	Std. diff.	Med. QQ	KS p-val[b]	Mean treated	Mean control	Std. diff.	Med. QQ	KS p-val[b]
Rainfall deviation	0.19	−1.34	0.88	0.32	0.00	0.92	−1.74	1.53	0.48	0.00
Population (log)	7.74	7.74	0.00	0.00	0.97	7.79	7.76	0.04	0.01	0.71
Percent African-American	0.28	0.28	0.01	0.01	0.65	0.28	0.27	0.06	0.02	0.39
Percentage with High School Degree	0.20	0.20	0.00	0.00	0.90	0.20	0.20	0.04	0.01	0.99
Percentage with College Degree	0.09	0.09	0.01	0.00	0.89	0.08	0.08	0.03	0.01	0.95
Median Household Income (log)	10.59	10.58	0.03	0.01	0.77	10.59	10.57	0.04	0.01	0.27
Percentage Below Poverty Line	0.13	0.14	0.02	0.01	0.94	0.14	0.14	0.01	0.01	0.80

[a] Match (I) performed without reverse caliper or sinks; (II) performed with reverse caliper set to $\varepsilon = 1.55$ and 1475 sinks.
[b] Kolmogorov-Smirnov p-values calculated from 5000 bootstrapped samples.

Table 5 presents the balance statistics for the Weak and Strong IV matches. Both matches are well balanced with small standardized differences and large KS test p-values. With the Strong IV match, we are able to achieve a much stronger instrument than we are able to with the national data. The standardized difference for the Weak IV match is 0.88, while the Strong IV match produces a standardized difference of 1.53. So we can nearly double the strength of the instrument in this application. Here, the Strong IV match uses 597 matched pairs instead of the 1335 matched pairs that are in the full sample. We now turn to estimating the effect of interest and a sensitivity analysis.

Table 6: Treatment effect estimates and sensitivity analysis for Georgia precincts match

	Weak IV			Strong IV		
	β	95% CI		β	95% CI	
	0.06	0.04	0.08	0.24	0.20	0.29

Γ	Weak IV	Strong IV
1.00	0.000	0.000
1.56	0.045	0.000
2.56	1.000	0.040
2.57	1.000	0.043
2.58	1.000	0.046
2.59	1.000	0.050
2.60	1.000	0.054

Table 6 contains the IV point estimate, the 95% confidence interval, and results from the sensitivity analysis. We first focus on the point estimates. When we match and do not strengthen the instrument, the IV point estimate is 0.06, which implies that an increase in turnout of one percentage point increases Democratic vote share by six-hundredths of a percent. The point estimate is statistically significant as the 95% confidence interval is bounded away from zero. We should note that the point estimate with the Georgia data is much smaller than we found with the non-Southern U.S. counties: 0.40 as compared to 0.06. This is likely due to the fact that state level confounders have been controlled for in this within state design. Next, we turn to the point estimate from the Strong IV design. In this application, the stronger instrument produces a much larger estimate than would be true with the Weak IV design. With point estimate now 0.24, the Strong IV estimate is 4 times larger than the Weak IV estimate. This point estimate is, however, still much smaller than the estimate based on counties.

We turn to a sensitivity analysis to understand whether the estimates are sensitive to departures from the assumption of as-if random assignment of the instrument. We expect the Strong IV design to produce an estimate that is less sensitive to hidden bias. We do find that the weaker instrument is more sensitive to biases from unmeasured confounders. With the Weak IV design $\Gamma = 1.56$ as compared to $\Gamma = 2.59$ with the stronger instrument. As such, in the Strong IV design, the odds of differing on the assignment of rainfall within matched pairs must differ by a factor of 2.59, while in the Weak IV design that factor is 1.56.

It is worth nothing that the Γ values with the Georgia data are much larger than in the analysis with counties, which implies that the smaller estimates from Georgia are less sensitive to hidden bias. As is often the case, a smaller more homogenous sample often produces estimates that are more robust to hidden bias (Rosenbaum 2005*a*, 2010). Thus, in terms of being robust to hidden bias, the smaller estimates from Georgia are more plausible than the county level analysis, which is subject to state level confounders.

3 Discussion and Conclusion

In many areas of political science natural experiments are the only way to make compelling causal claims. The difficulty is that analysts often mistake haphazard treatment assignment for as-if random assignment. Haphazard treatment assignment, while preferable to self-selection by units, still requires considerable care to ensure that causal claims are credible. Instrumental variables are a valuable tool in the analysis of natural experiments, but IV presents a number of well known pitfalls. Here, we highlight a less well known interaction among the the IV assumptions. When treatment assignment is haphazard as it is in a natural experiment, imbalance on covariates can occur. Such imbalances, even when minor, combined with a weak instrument, produce inconsistent estimates (Small and Rosenbaum 2008).

We have outlined a strategy based on matching to cope with the challenges that arise when instruments are weak and imbalances are present. We use matching to make units similar in terms of observed pre-instrument covariates, while at the same time we find matched pairs that are as dissimiliar as possible on the encouragement from the instrument. The end result are matched pairs where imbalances are removed and the instrument is as strong as possible. The matched design also allows us to easily apply Rosenbaum's (2002*b*) method of sensitivity analysis. In general, we think this nonparametric approach to IV is superior to the more conventional methods of estimation for IV like two-stage least squares. In our application, we removed imbalances that were in the original data and created a subset of matched pairs where the instrument was stronger. Our method produces both larger point estimates and were less sensitive to hidden bias. Our analysis also improves on the original by using a within state design. This design produces estimates that are much smaller than those based on a national sample.

Finally, we think it is worth concluding with some more general remarks about IV. Applications of IV can be done within the potential outcomes framework or it can be done within a more traditional econometric approach. The danger of divorcing IV from the potential outcomes framework is that the analyst can lose track of what is being estimated: the local average treatment effect. The research question in this case is a general one about whether the composition of the electorate that is marginal for voting tends to be more Democratic. Does IV provide an answer to this general question? No. Instead IV allows us to observe that for those places where turnout was lower because of unusual rainfall Democratic vote share was lower. Thus the answer to this general question is based on a sub-population whose political participation hinges on weather patterns. Is this a sub-population that is relevant to political science theory? We do not attempt to offer an answer here, but analyses based on IV need to consider such questions. A more general answer to this research question requires that we find other interventions that produce haphazard reductions in turnout. If later research finds other instruments that lower turnout and Democratic vote share still declines, then a more general causal pattern will have been established. In sum, successful use of IV often provides a limited answer to more general questions. Divorcing IV from the potential outcomes framework makes such confusion more likely.

References

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434):444–455.

Baiocchi, Mike, Dylan S. Small, Scott Lorch and Paul R. Rosenbaum. 2010. "Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants." *Journal of the American Statistical Association* 105(492):1285–1296.

Betz, Timm. 2013. "Robust Estimation with Nonrandom Measurement Error and Weak Instruments." *Political Analysis* 21(1):89–96.

Bound, J., D.A. Jaeger and R.M. Baker. 1995. "Problems with Intrustmental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90(430):443–450.

Cochran, William G. and Donald B. Rubin. 1973. "Controlling Bias in Observational Studies." *Sankyha* 35:417–446.

Dunning, Thad. 2009. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61(2):282–293.

Dunning, Thad. 2010. Design-Based Inference: Beyond the Pitfalls of Regression. In *Rethinking Social Inquiry*, ed. Henry E. Brady and David Collier. 2nd ed. Lanham, MD: Rowman and Littlefield Publishers pp. 273–312.

Fraga, Bernard and Eitan Hersh. 2010. "Voting Costs and Voter Turnout in Competitive Elections." *Quarterly Journal of Political Science* 5(4):339–356.

Gomez, Brad T., Thomas G. Hansford and George A. Krause. 2007. "The Republicans Should Pray for Rain: Weather Turnout, and Voting in U.S. Presidential Elections." *Journal of Politics* 69(3):649–663.

Greevy, Robert, Bo Lu, Jeffery H. Silber and Paul Rosenbaum. 2004. "Optimal Multivariate Matching Before Randomization." *Biostatistics* 5(2):263–275.

Hansford, Thomas G. and Brad T. Gomez. 2010. "Estimating the Electoral Effects of Voter Turnout." *American Political Science Review* 104(2):268–288.

Hodges, J. L. and E.L. Lehmann. 1963. "Estimates of Location Based on Ranks." *The Annals of Mathematical Statistics* 34(2):598–611.

Holland, Paul W. 1988. "Causal Inference, Path Analysis, and Recursive Structural Equation Models." *Sociological Methodology* 18:449–484.

Imbens, Guido W. and Paul R. Rosenbaum. 2005. "Robust, Accurate Confidence Intervals With a Weak Instrument: Quarter of Birth and Education." *Journal of The Royal Statistical Society Series A* 168(1):109–125.

Keele, Luke J. and William Minozzi. 2012. "How Much is Minnesota Like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data." *Political Analysis* Forthcoming.

Lu, Bo, Elaine Zutto, Robert Hornik and Paul R. Rosenbaum. 2001. "Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse." *Journal of the American Statistical Association* 96(456):1245–1253.

Lu, Bo, Robert Greevy, Xu X. and Beck C. 2011. "Optimal Nonbipartite Matching and its Statistical Applications." *The American Statistician* 65(1):21–30.

Neyman, Jerzy. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5(4):465–472. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).

Rosenbaum, Paul R. 1996. "Identification of Causal Effects Using Instrumental Variables: Comment." *Journal of the American Statistical Association* 91(434):465–468.

Rosenbaum, Paul R. 2002*a*. "Covariance Adjustment In Randomized Experiments and Observational Studies." *Statistical Science* 17(3):286–387.

Rosenbaum, Paul R. 2002*b*. *Observational Studies*. 2nd ed. New York, NY: Springer.

Rosenbaum, Paul R. 2005*a*. "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Obervational Studies." *The American Statistician* 59(2):147–152.

Rosenbaum, Paul R. 2005*b*. Observational Study. In *Encyclopedia of Statistics in Behavioral Science*, ed. Brian S. Everitt and David C. Howell. Vol. 3 John Wiley and Sons pp. 1451 – 1462.

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

Rosenzweig, Mark R. and Kenneth I. Wolpin. 2000. "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature* 38(4):827–74.

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

Sekhon, Jasjeet S. 2011. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package For R." *Journal of Statistical Software* 42(7):1–52.

Sekhon, Jasjeet S. and Rocío Titiunik. 2012. "When Natural Experiments are Neither Natural Nor Experiments." *American Political Science Review* 106(1):35–57.

Small, Dylan and Paul R. Rosenbaum. 2008. "War and Wages: The Strength of Instrumental Variables and Their Sensitivity to Unobserved Biases." *Journal of the American Statistical Association* 103(483):924–933.

Sovey, J. Allison and Donald P. Green. 2011. "Instrumental Variables Estimation in Political Science: A Readers' Guide." *American Journal of Political Science* 55(1):188–200.

Staiger, D. and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65:557–586.

Wald, A. 1940. "The Fitting of Straight Lines if Both Variables Are Subject to Error." *The Annals of Mathematical Statistics* 11:284–300.