Submitted to the Annals of Applied Statistics

HOW STRONG IS STRONG ENOUGH? STRENGTHENING INSTRUMENTS THROUGH MATCHING AND WEAK INSTRUMENT TESTS[*]

By Luke Keele[†] and Jason W. Morgan[‡]

Penn State University[†] *and The Ohio State University*[‡]

In a natural experiment, treatment assignments are made through a haphazard process that is thought to be as-if random. In one form of natural experiment, encouragement to accept treatment rather than treatments themselves are assigned in this haphazard process. This encouragement to accept treatment is often referred to as an instrument. Instruments can be characterized by different levels of strength depending on the amount of encouragement. Weak instruments that provide little encouragement may produce biased inferences, particularly when assignment of the instrument is not strictly randomized. A specialized matching algorithm can be used to strengthen instruments by selecting a subset of matched pairs where encouragement is strongest. We demonstrate how weak instrument tests can guide the matching process to ensure that the instrument has been sufficiently strengthened. Specifically, we combine a matching algorithm for strengthening instruments and weak instrument tests in the context of a study of whether turnout influences party vote share in US elections. It is thought that when turnout is higher, Democratic candidates will receive a higher vote share. Using excess rainfall as an instrument, we hope to observe an instance where unusually wet weather produces lower turnout in an as-if random fashion. Consistent with statistical theory, we find that strengthening the instrument reduces sensitivity to bias from an unobserved confounder.

1. Gifts of nature.

1.1. *Natural experiments.* In the social sciences, analysts are often interested in the study of causal effects, but in many contexts randomized experiments are infeasible. When this is the case, 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" as a strategy for estimating unbiased casual

[*] We thank Hyunseung Kang, Mike Baiocchi, Dylan Small, Paul Rosenbaum, Wendy Tam Cho, Jake Bowers, Danny Hidalgo, and Rocío Titiunik for helpful comments and discussion.

Keywords and phrases: causal inference, instrumental variables, matching, weak instruments

effects (Rosenzweig and Wolpin 2000, pg. 872). Many view natural experiments as a close second best to a true randomized experiment (Angrist and Pischke 2010; Dunning 2012).

In a natural experiment, some units either obtain or are denied treatment in a haphazard manner. The general hope is that nature has reduced biases that may interfere with our ability to observe causal effects. The difficulty, of course, is that haphazard assignment to treatment may be a far cry from a randomized experiment, where randomization is a known fact. As a result, many natural experiments require more complex forms of statistical adjustment than would be necessary for a randomized experiment. For example, matching methods are often used to increase comparability across treated and control groups in a natural experiment (Baiocchi et al. 2010; Keele, Titiunik and Zubizarreta 2014; Zubizarreta, Small and Rosenbaum 2014). It is such adjustments that render natural experiments closer in form to observational studies. In some instances, however, we may wish to aid haphazard assignment in a different way. That is, we may wish to find units that are more disparate than naturally rendered by circumstance. Here we consider one of those instances.

1.2. *A natural experiment studying the effect of turnout on vote share.* It is often assumed that many of the Democratic party's natural constituencies are less likely to vote on election day. That is, younger voters, minorities, and citizens with lower levels of income often vote less frequently (Wolfinger and Rosenstone 1980; Nagler 1991; Keele and Minozzi 2012). The logical conclusion to the evidence that these groups tend to vote Democratic is that higher levels of voter turnout should result in increased vote share for Democratic candidates (Hansford and Gomez 2010). One major difficulty with evaluating this proposition is that there may be common causes for both voter turnout and vote share. If such common causes are unobservable, we should be hesitant to draw causal inferences about turnout and vote share. Alternatively, one strategy is to determine whether there is some source of variability in turnout that does not reflect individual choices about voting but is, instead, by chance. Although the choice to vote on election day is determined by many factors such as interest in politics, exposure to mobilization efforts, and socio-economic status, there may be a factor that could haphazardly encourage or discourage participation on election day.

Here, we focus on a haphazard contrast first used by Hansford and Gomez (2010). That is, bad weather may serve as a haphazard disincentive to voting. While civic duty and political interest may induce political participation, for many voters a soggy day may be enough to dissuade a trip to the voting booth (Gomez, Hansford and Krause 2007). That is, rainfall, specifically unusually wet weather or excess rainfall, serves as a haphazard nudge to not vote. In the language of research designs, rain may serve as an instrument. An instrument is a random nudge to accept a treatment. This nudge to accept treatment may or may not induce acceptance of the treatment, and the nudge can affect the outcome only through the treatment. In our study, we seek to compare locations that have similar observable characteristics, but where one location had unusually wet weather on election day. In our design, rain serves as random nudge against voting that we assume can only affect the outcome, vote share, through turnout.

What question can we answer using an instrument as a natural experiment? Conditional on a set of identification assumptions, we seek to estimate the causal effect of turnout on vote share for the subset of places 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. As such, our estimand only refers to those places that respond to the rainfall instrument: the places that are sensitive to unusual rainfall patterns on election day.

We use the excess rainfall instrument to illustrate how an instrumental variables analysis can be conducted using matching algorithms; specifically matching methods that can strengthen an instrumental variable (Baiocchi et al. 2010; Zubizarreta et al. 2013). We extend those matching methods by pairing them with tests for weak instruments from the economics literature (Stock and Yogo 2005). We demonstrate how weak instrument tests can be used to aid the matching process such that the analyst can know whether a strengthened instrument is strong enough or whether the matching may need further refinement.

1.3. *Review of key concepts: instrumental variables and weak instruments.* As we noted above, an instrument is a nudge to accept treatment. As applied to natural experiments, an instrument, such as rainfall, is meant to mimic the randomized encouragement design (Holland 1988). 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. Subject to a set of causal identification assumptions, the method of instrumental variables can be used to estimate the effect of the treatment as opposed to the effect of the encouragement. The causal effect identified by an instrument is often referred to as local average treatment effect (LATE) or complier average causal effect (CACE) (Imbens and Angrist 1994).

Identification of the IV causal effect requires five assumptions as outlined by Angrist, Imbens and Rubin (1996). One of these assumptions, the exclusion restriction, receives considerable attention. Under this assumption, we must assume that the instrument has no direct effect on the outcome. In our application, the exclusion restriction implies that excess rainfall affects vote share only by reducing turnout on election day. To violate the exclusion restriction, excess rainfall must influence vote share directly. That is, there must be some aspect of precipitation patterns that change partisan preferences in an election, which seems unlikely. The use of rainfall deviations further bolsters the case for the exclusion restriction, since even if weather patterns did directly affect vote preferences, it seems less likely that a haphazard deviation from normal weather alters voter preferences in any significant way. Thus, while we cannot verify that the exclusion restriction holds, it appears to be plausible in this application.

While the exclusion restriction requires careful evaluation, two of the other IV assumptions are often unlikely to hold exactly when the instrument is haphazardly assigned, as would be the case with rainfall patterns on election day. One of these assumptions is that the assignment of the instrument must be ignorable, or as-if random. 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 by design. In natural experiments, it is often unclear that instrument ignorability holds since assignment to encouragement happens through some natural, haphazard process and is not a controlled, probabilistic assignment mechanism. For any natural experiment, the possibility always remains that the instrument is not as-if randomly assigned. Analysts can use a sensitivity analysis to observe whether study conclusions are sensitive to this assumption (Rosenbaum 2010, 2002a, ch. 5).

Additionally, the instrument must have a nonzero effect on the treatment. However, even when that effect is nonzero, instruments may be weak. An instrument is said to be weak if manipulation of the instrument has little effect on treatment (Staiger and Stock 1997). When the instrument has a weak effect on the treatment, poor coverage of confidence intervals can result. In fact, the most common method of estimation used with instrumental variables, two-stage least squares (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 such effects are nonexistent. In our application, while rainfall explains some variation in turnout, it explains a fairly small portion of that variation. Note in our application there is single weak instrument, which is a distinct problem from the case with many weak instruments. When there are many weak instruments, 2SLS produces biased point estimates as well as standard errors that are too small (Staiger and Stock 1997). See Chamberlain and Imbens (2004); Chao and Swanson (2005) for statistical models for many weak instruments.

In short, prima facie, we might suspect that rainfall is not as-if randomly assigned on election day, and may be a weak instrument. However, for an instrument like rainfall we cannot consider the problem of ignorable instrument status as separate from the difficulties caused by weak instruments. Small and Rosenbaum (2008) show that when an instrument is weak even small departures from ignorability of instrument assignment status produces bias even in large samples. Small and Rosenbaum (2008) also prove that a strong instrument is more robust to departures from ignorability even in smaller sample sizes. Thus they show that if ignorability does not hold, a smaller study with a stronger instrument will be less sensitive to bias than a weak instrument used in a much larger study. Below we outline and extend a matching method designed to combat the difficulties that arise when instruments are weak and not as-if randomly assigned.

1.4. *Data: covariates and measurement.* The data describe vote share and turnout at the county level for the 2000 US presidential election. Turnout is measured as a percentage of votes cast for presidential candidates divided by the voting age population, while vote share is measured as the percentage of the the two-party vote share received by the Democratic presidential candidate. Overall, the data set includes more than 1900 counties across 36 US states.[1] The year 2000 is hospitable to our project, since across the country, there was a large variation in rainfall on election day. In several other presidential election years, there was little rainfall in most places across the US. For the rainfall instrument, we use the covariate developed in the original analysis, which is measured as excess rainfall (Gomez, Hansford and Krause 2007). It is the difference between the amount of rainfall recorded on election day and the average rainfall in the period around election day. Thus positive (negative) values indicate greater (lesser) than average rainfall. So under this design, we examine whether unusually rainy weather discourages turnout.

In our analysis, we added several covariates that are likely to be related to turnout and electoral outcomes based on past research in political science (Wolfinger and Rosenstone 1980; Nagler 1991). These covariates include the

[1] As in the original paper, we also exclude Southern counties from the data, since historically turnout in the South is affected significantly by more restrictive voting requirements.

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. These covariates should be balanced if the rainfall instrument is as-if randomly assigned. 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 to avoid bias from conditioning on a concomitant variable (Rosenbaum 1984).

1.5. *Outline of the paper.* In Section 2 we review optimal nonbipartite matching and how it may be used to strengthen instruments such as rainfall. This form of matching allows us to form matched pairs that are close as measured by covariates, but are distant in terms of the amount of excess rainfall on election day. We also introduce a new search criterion for the matching based on weak instrument tests. The proposed search criterion allows analysts to readily understand whether a proposed match has sufficiently strengthened the instrument. Sections 3 and 4 present results from our case study. We report results from three matches in Section 3. In Section 4, we estimate causal effects and use a sensitivity analysis to assess whether hidden bias might alter our inference under the stated assumptions. Section 5 concludes.

2. Optimal nonbipartite matching to control for overt biases and strengthen instruments.

2.1. *Notation.* First, we introduce notation for the paired randomized encouragement design(Rosenbaum 1996, 2002b). It is this experimental design that IV with matching mimics. There are I matched pairs, $i = 1, \dots, I$, and the units within matched pairs are denoted with $j \in \{1, 2\}$. We form these pairs by matching on observed covariates, $\mathbf{x}_{ij}$, which are measured before assignment to the instrument. Let W_{ij} denote the value of the instrument, excess rainfall, for county j in possible pair i. We wish to form matched pairs, such that $W_{i1} > W_{i2}$ in each pair. That is within matched pairs, one county is discouraged to accept a higher dose of treatment (turnout) by a higher level of excess rainfall. We denote the county with a higher value for W_{ij} as $Z_{ij} = 1$, and the other county with a smaller value for W_{ij} is denoted by $Z_{ij} = 0$, so that $Z_{i1} + Z_{i2} = 1$ for $i = 1, \dots, I$. If $Z_{ij} = 1$ the unit is discouraged to a greater extent and county ij receives treatment at dose

d_{Tij}, and if $Z_{ij} = 0$ county ij receives treatment at dose d_{Cij}, where the subscript T denotes treatment and C denotes control.

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, $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$.

2.2. *Matching to increase instrument strength.* A natural experiment may produce an instrument that is characterized as weak. 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 when most units ignore the encouragement to take the treatment. With a continuous instrument, such as excess rainfall, we might imagine that the ideal matched pair of subjects ik and il would have $\mathbf{x}_{ik} = \mathbf{x}_{il}$ but the difference, $W_{ik} - W_{il}$, would 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 creates comparable units with a large difference in terms of encouragement allowing for a stronger instrument. How might we implement such a match?

Baiocchi et al. (2010) demonstrate how to use nonbipartite matching with penalties to implement this ideal IV match. 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 minimize distances on observables within matched pairs subject to a penalty on instrument distance as measured by $W_{i1} - W_{i2}$, the distance between the observations in the matched pair on the instrument. The distance penalty, p, is defined as

$$p = \begin{cases} (W_{i1} - W_{i2})^2 \times c & \text{if } W_{i1} - W_{i2} < \Lambda \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where Λ is a threshold defined by the analyst. Note that the scale for Λ depends on the metric for W_{ij}. The penalty, p, is defined such that a smaller value of $W_{i1} - W_{i2}$ 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. Matched distances on the instrument less than Λ receive larger penalties and thus are less likely to be matched. The result is that units that are alike on observables but more different on the instrument tend to be matched.

To be fully effective, however, the penalized matching generally must be combined with "sinks" (Lu et al. 2001). Matching with penalties alone tends to produce some matched pairs that are distant on the instrument but have suboptimal within matched pair distances on observables. That it, strengthening the instrument often makes covariate balance worse. To improve balance, we use sinks to discard the observations that are hardest to match well. 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 worst matches. Thus the optimal 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 (Baiocchi et al. 2010). 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). Both Lorch et al. (2012) and Baiocchi et al. (2012) present examples of this type of near-far match in medical applications. Zubizarreta et al. (2013) shows how one can implement a near-far match with integer programming, so the analyst can impose different types of balance constraints on different covariates, while also strengthening the instrument.

2.3. *A new search criterion for stronger instrument matching.* Through the use of penalties, we can form a set of matched pairs where the instrument strength, as represented by the within pair distance on the instrument, is larger than occurs without penalties. One question remains: how strong does the instrument need to be? The matching process increases relative instrument strength, but it is silent on how large we should seek to make the difference on the instrument within matched pairs. What would be useful is some absolute standard of instrument strength that we might use to select the matching parameters.

The literature on weak instrument tests in econometrics suggests an absolute standard that we might employ to guide the matching. First, we review weak instrument tests, and then we incorporate a weak instrument test into the matching process. Stock and Yogo (2005) suggest the following test for a weak instrument based on

$$D_i = \gamma W_i + \nu_i \tag{2}$$

which in our application is turnout, D_i, regressed on excess rainfall, W_i, and we assume ν_i is independent of W_i, since the instrument is as-if randomly assigned. The concentration parameter is a unitless measure of instrument strength and is defined as:

$$\mu^2 = \frac{\gamma W_i W_i \gamma}{\sigma^2_\nu}$$

where, σ^2_ν, is the variance of the residual term in Equation 2. Stock and Yogo (2005) use the F-statistic for testing the hypothesis that $\gamma = 0$ as an estimator for μ^2. However, using F to test the hypothesis of nonidentification ($\gamma = 0$) is not a conservative enough test for the problems caused by weak instruments. Stock and Yogo (2005) recommend using F to test the null hypothesis that μ^2 is less than or equal to the weak instrument threshold, against the alternative that it exceeds the threshold. When there is a single instrument, the first-stage F statistic must generally exceed 10 for the instrument to be sufficiently strong (Stock and Yogo 2005).

In the context of strengthening instruments, we expect μ^2, as measured through the F-test, to increase as Λ gets larger. This suggests that the weak instrument test contains useful information about how to select Λ. When matching to strengthen the instrument, the analysts must select both Λ, the threshold at which penalties are enforced based on instrument strength, and e, the number of sinks. Each combination of these two parameters produce a match with a specific instrument strength and sample size. The analyst must then select one match based on these two parameters as a final match. We augment this part of the design process with the weak instrument test.

First, we treat the matching process as a grid search over both Λ and e. For each combination of Λ and e, we perform the weak instrument test by estimating a regression of D_i on W_i using the matched data without any other covariates. From this regression model, we record the F-test statistic and the R^2. We then use a surface plot to examine instrument strength for combinations of Λ and e. The plot allows the analysts to clearly see which combination of Λ and e produce a match where the instrument may be deemed sufficiently strong. We can also add a contour line to the plots at the point where $F \geq 10$ to demarcate the region where the combination of e and Λ produce a match where the instrument is sufficiently strong. While it is true that these F-statistics will be highly correlated especially around the line demarcating those matches that are weak and strong as characterized by the test, we still think the information is useful. While we could correct for multiple testing, this threshold serves as a useful heuristic and in general, analysts should pick a match well beyond the $F = 10$ threshold, especially when instrument assignment does not appear to be strictly as-if random. As an additional guide, we can create a similar plot that records the R^2 from the weak instrument test regression.

3. The nonbipartite match.

3.1. *Covariate balance before matching.* Before matching, we assess whether excess rainfall follows an as-if random assignment pattern. If excess rainfall is a valid instrument, we should expect that characteristics like levels of education and income to be balanced between counties with normal and those with unusual amounts of rainfall on election day. 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 a series of balance tests on the county level covariates. In Table 1 for each covariate we report means, the absolute standardized difference in means (the absolute value of the difference in means divided by the standard deviation before matching), and the p-value from the Kolmogorov-Smirov (KS) test.

We find that rainfall patterns in 2000 were not as-if random. For the 2000 election cycle, counties with excess rainfall were less likely to be Hispanic and African American and had lower levels of income and education. A general rule of thumb is that matched standardized differences should be less than 0.20 and preferably 0.10 (Rosenbaum 2010). With one exception, all the standard differences exceed 0.10, with several being above 0.30. Moreover, even when the centers of the distributions appear to be similar, as is true for the percentage of residents that are below the poverty line, the KS test indicates that other moments of the distribution differ as every KS test p-value is less than 0.01. In summary, the balance test results clearly demonstrate that the haphazard nature of excess rainfall does not produce the same level of balance that would be produced by randomization. Coupled with the fact that rainfall is a relatively weak instrument for turnout—the correlation between rainfall and turnout is 0.08 in 2000—it would be dangerous to make

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 black	0.02	0.03	0.18	0.00
Percent Hispanic	0.03	0.07	0.42	0.00
High school educ.	0.80	0.82	0.24	0.00
College educ.	0.16	0.20	0.59	0.00
Median HH income (log)	10.45	10.52	0.30	0.00
Poverty rate	0.13	0.13	0.00	0.00

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

inferences based on the excess rainfall instrument without further statistical adjustments.

3.2. *How the matching was done.* In the matching, we calculated the pairwise distances between the counties included in the sample. We used a rank-based Mahalanobis distance metric, which is robust to highly skewed variables (Rosenbaum 2010). We also applied a large penalty for geographic contiguity so that the algorithm avoids matching contiguous counties, if possible, subject to minimizing imbalances on the covariates. The logic behind this adjacency constraint is as follows. Take units A and B which are adjacent. Unit A records 1 inch of rain above average. Unit B records .25 inches of rain above average. However, the weather station in B is far from the border with A and with additional weather stations we would record .5 inches of rain in unit B. If we pair units A and B, we over-estimate the discouragement from rainfall by recording $z_{ij} - z_{ik} = .75$ instead of the actual discrepancy of $z_{ij} - z_{ik} = .5$. Take unit C which is non-adjacent to A. For units A and C, we record $z_{ij} - z_{ik} = .75$, this discrepancy is much less likely be a function of measurement error due to adjacency.

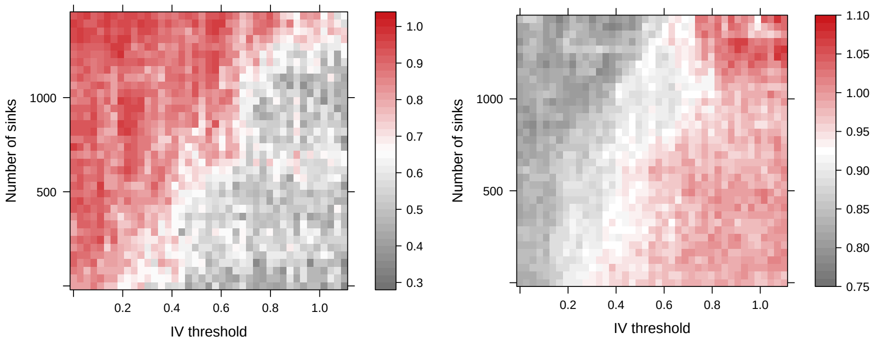
We would also expect that rainfall patterns are spatially correlated. In our context, such spatial correlation could result in a violation of the no interference component of the stable unit treatment value assumption (SUTVA) (Rubin 1986). SUTVA is assumed under the potential outcomes approach to causal inference, and is assumed in the IV framework of Angrist, Imbens and Rubin (1996). A spatially correlated instrument may not induce a a violation of SUTVA, but a spatially correlated instrument would require adjusting our estimates of uncertainty. How might a spatially correlated instrument violate SUTVA? Interference could occur if rainfall recorded in one county decreases turnout in a nearby county, but the nearby county records a small amount of excess rainfall. Our approach to matching counties that are not adjacent is also an attempt to bolster the plausibility of SUTVA. To that end, we characterize interference as partial interference (Sobel 2006) where observations in one group may interfere with one another but observations in distant places do not. Thus by not matching adjacent counties we hope to reduce the likelihood that rainfall recorded on election day in one location is less likely to travel to a location farther away. See Zigler, Dominici and Wang (2012) for another example where spatial correlation is characterized as a possible SUTVA violation.[2]

As we proposed above, we perform a grid search over combinations of Λ and e. In this case, we used a grid search for values of Λ from 0 to 1.10 and sinks between 0 and 1435. We selected the maximum value for Λ as the value equal to four times the standard deviation on the unmatched pairwise differences in excess rainfall. We set the maximum number of sinks to 1435, which means dropping nearly 80% of the sample. Overall, this implies 41 different values for Λ and 36 values for the sinks, producing 1476 different match specifications.

We deemed 1476 different matches a sufficient number of matches to explore the matching space, though this decision was based on informal reasoning rather than any formal calculations. For each specification, optimal nonbipartite matching was performed and balance statistics on the covariates were recorded. We also recorded the strength of the instrument as recorded by the standardized difference on the excess rainfall measure within matched county pairs and the results from a weak instrument test.

3.3. *Mapping the matching space and an initial match.* We summarize the results of all these matches using a set of figures that summarize the matching space. Figure 1 contains two plots that summarize the matches for combinations of sinks and values of Λ. In Figure 1a, we summarize balance by plotting the mean p-value from the KS-test for each possible match. The match at the origin in the plot (bottom left corner) is the match with no sinks and Λ set to zero. This represents a standard matching that uses the entire population of counties and places no penalties on the within-pair difference in excess rainfall, and therefore does not strengthen the instrument. In the plot, we observe that as we strengthen the instrument by increasing the

[2] However, we also found that allowing contiguous counties to be matched did not alter our inferences.



(a) Mean KS-test p-values

(b) Std. Diff on Excess Rainfall

Fig 1: In panel (1a) we record the mean KS test p-value for each match given a combination of sinks and a value for Λ the penalty for strengthening the instrument. In panel (1b), we record the absolute standardized difference within matched counties on the excess rainfall measure. Each square in the plots represents either an average p-value or the standardized difference on excess rainfall for each of the 1476 matches.

value of Λ we tend to make balance worse. However, if we drop observations by adding sinks, we can increase the strength of the instrument and preserve balance.

In Figure 1b, we summarize the match by plotting the absolute standardized difference within matched pairs on the measure of excess rainfall, a measure of instrument strength. Clearly increasing the value of Λ increases the strength of the instrument as measured by the standardized difference. As we might expect, instrument strength is somewhat invariant to the number of sinks. That is, if we do little to strengthen the instrument adding sinks does little to further strengthen the instrument. However, the instrument is strongest when we set a large penalty on Λ and use many sinks.

Examination of both plots begs the question of which match we might prefer. There appear to be a number of acceptable matches in terms of those where the instrument is stronger due to the penalties, while balance remains acceptable. For example, take one match where $\Lambda = 0.55$, which is equal to two standard deviations on the pairwise differences in rainfall with 675 sinks, which drops approximately one third of the observations. This match is comprised of 625 matched county pairs instead of the full set of

TABLE 2
Balance statistics for US county matches

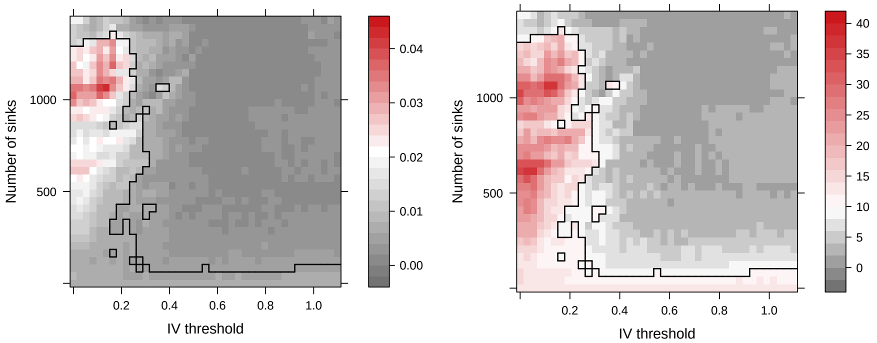
	Medium IV[a] $I = 625$ matched pairs			
	Mean treated	Mean control	Std. diff.	KS p-val[b]
Rainfall deviation	0.35	0.06	0.96	0.00
Population (log)	10.04	10.07	0.02	0.70
Percent African-American	0.02	0.02	0.01	0.58
Percent Hispanic	0.05	0.04	0.01	0.68
High school educ.	0.81	0.81	0.03	0.83
College educ.	0.17	0.17	0.00	0.99
Median Household income (log)	10.46	10.46	0.02	0.78
Poverty rate	0.13	0.13	0.04	0.64

[a] Match performed with $\varepsilon = 0.55$ and 675 sinks.
[b] Kolmogorov-Smirnov p-values calculated from 5000 bootstrapped samples.

962 matched county pairs available if we don't use any sinks. For this match, the standardized difference on the rainfall measure increases from 0.82 to nearly 1.0. For this number of sinks, we found that stronger instruments produced levels of balance we thought acceptable. Table 2 contains the balance statistics for this match. For this match, all standardized differences are less than 0.10, and the smallest p-value from the KS test is 0.58. As such, this would appear to be a successful match, in that we have increased the strength the instrument, maintained balance, and not discarded a high number of observations.

Next, we examine the results from a weak instrument test for this match. For this match, the R^2 is 0.00458 and the value from the F-test is well below the standard threshold of 10. It is worth noting that in simulation evidence, 2SLS confidence intervals had poor coverage properties when the first stage R^2 fell below 0.05 (**?**). This also highlights a difficulty with selecting an acceptable match. An increase in the standardized difference on the instrument may not translate into sufficiently strong instrument. We next use weak instrument tests to guide the selection of a final match.

3.4. *Selection of a match with weak instrument tests.* For the 1476 total matches we performed, we also recorded information from weak instrument tests. In the panels of Figure 2 we plot quantities from the weak instrument tests for each match. In Figure 2a, we plot the R^2 and Figure 2b contains



(a) R^2 for each match.

(b) F-statistic for each match.

Fig 2: In panel (2a) we record the R^2 for each match given a combination of sinks and a value for Λ the penalty for strengthening the instrument. In Figure 2b, we record the F-statistic from a regression of turnout on excess rainfall using the matched data. The dark line in both panels demarcates matches where the F-statistic is greater than 10. Matches to the left of the line pass the weak instrument test.

the F-test statistic for each match. Each panel also contains a contour line that demarcates the region where the matching produces an F-test statistic larger than 10. We can clearly see that a minority of the matches produce results that pass the weak instrument test. In fact, we observe that for a value of Λ above 0.20 we rarely pass the weak instrument test unless we use very few sinks, and the strongest match uses a large number of sinks but cannot strengthen the match much above a value of 0.20 for Λ.

We selected as a final match, the match the produced the largest F-statistic and R^2 value in Figure 2. In Table 3 we present balance statistic results for this match as well as the match utilizes the full sample and does not strengthen the instrument. In general, there is relatively small amount that we can strengthen the instrument in this example in terms of the standardized difference. For other covariates, both matches produce acceptable levels of balance as all the standardized differences are less than 0.10 and the smallest KS test p-value is 0.40.

This example illustrates the value of our approach. Without the results from the F-test, we suspect most analysts would have selected a match much like the one we reported in Table 2. In that match, we produced a higher

Table 3

Balance statistics for two matches. The standard IV match uses the full sample and does not strengthen the instrument. The Strong IV match is based on the match that produces the largest F-statistic from the weak instrument test.

	(I) Standard IV[a] Match $I = 962$ matched pairs				
	Mean treated	Mean control	Std. diff.	Med. QQ	KS p-val[b]
Rainfall deviation	0.32	0.07	0.82	0.27	0.00
Population (log)	10.18	10.22	0.03	0.01	0.40
Percent African-American	0.02	0.02	0.00	0.01	0.97
Percent Hispanic	0.05	0.05	0.03	0.01	0.76
High school educ.	0.81	0.81	0.01	0.01	1.00
College educ.	0.17	0.17	0.04	0.01	0.83
Median Household income (log)	10.48	10.48	0.01	0.01	0.86
Poverty rate	0.13	0.13	0.02	0.01	0.76

	(II) Strong IV[a] Match $I = 429$ matched pairs				
	Mean treated	Mean control	Std. diff.	Med. QQ	KS p-val[b]
Rainfall deviation	0.29	0.04	0.84	0.28	0.00
Population (log)	10.50	10.55	0.03	0.01	0.77
Percent African-American	0.03	0.03	0.00	0.01	0.99
Percent Hispanic	0.04	0.04	0.05	0.01	0.95
High school educ.	0.81	0.81	0.00	0.01	1.00
College educ.	0.18	0.18	0.06	0.01	0.88
Median Household income (log)	10.49	10.49	0.01	0.01	0.68
Poverty rate	0.13	0.13	0.02	0.01	0.88

[a] Match (I) performed without reverse caliper or sinks; (II) performed with $\varepsilon = 0.1375$ and 1066 sinks.

[b] Kolmogorov-Smirnov p-values calculated from 5000 bootstrapped samples.

standardized difference on the instrument, but the match failed the weak instrument test, as did most of the matches we produced. Our example also demonstrates that while instruments can be strengthened through matching, there may be limits to amount of strengthening that can occur.

In our application to find a match where the instrument has been sufficiently strengthened, we had to focus on a much smaller set of counties. In the standard match, there are 962 matched pairs, so 962 counties have been encouraged to have lower turnout by rainfall. However, in the strong IV match only 429 counties are encouraged to have lower turnout. As such, we have altered the estimand through matching since we deem a much smaller fraction of the study population to be compliers. However, we can explore whether the larger set of compliers differs from the set of compliers in the strong IV match in terms of observed characteristics. Table 4 contains mean differences between the counties that are compliers in the standard match but not included in the strong IV match and those counties that are treated as compliers in the strong IV match. In terms of observable characteristics, these two sets of counties are quite similar. This implies that there are simply not enough counties with low amounts of excess rainfall available for matching to allow for a large difference in rainfall in all the matched pairs.

Table 4

Comparison of compliers between standard IV match and strong IV match. This comparison is between the set of counties with higher rainfall in the strong IV match, and the non-overlapping set of counties with higher rainfall in the standard match.

	Standard IV Match	Strong IV Match
	Mean	Mean
Rainfall deviation	0.31	0.29
Population (log)	10.00	10.50
Percent African-American	0.02	0.03
Percent Hispanic	0.05	0.04
High school educ.	0.81	0.81
College educ.	0.17	0.18
Median Household income (log)	10.46	10.49
Poverty rate	0.13	0.13

4. Inference, estimates, and sensitivity analysis.

4.1. *Randomization Inference for an Instrument.* Next, we outline IV testing and estimation via randomization inference. Following Rosenbaum (1999), we assume that the effect of encouragement on response is proportional to its effect on the treatment dose received,

$$r_{Tij} - r_{Cij} = \beta(d_{Tij} - d_{Cij}). \tag{3}$$

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

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

Under this model of effects, the response will take the same value regardless of the value of Z_{ij}, which makes this model of effects consistent with 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 the 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}. In this model of effects, the treatment effect varies from unit to unit based on the level of D_i as measured by $(d_{Tij} - d_{Cij})$. If the unit received no dose, then $r_{Tij} - r_{Cij} = \beta(d_{Tij} - d_{Cij}) = 0$. In the application, a full dose occurs when $D_{ij} = 100$. For units that take the full dose, the effect is $\beta \times 100$, while a unit with half a dose would have an effect $\beta \times 50$.

Given this model of effects, we wish to test whether the treatment is without effect, estimate a point estimate, and form a confidence interval. 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 inferences about β using the observed quantity $R_{ij} - \beta_0 D_{ij} = U_{ij}$ 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 $|U_{ij}|$ from 1 to I. We calculate Wilcoxon's signed rank statistic, U, as the sum of the ranks for which $U_{ij} > 0$. 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 U 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 that $\beta_0 = 0$.

A point estimate for β is obtained using the method of Hodges and Lehmann (1963). The Hodges-Lehmann 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 U equals $I(I+1)/4$ when U is computed from $R_{ij} - \beta_0 D_{ij}$. A 95% confidence interval for the treatment effect is formed by testing a series of

TABLE 5
Treatment effect and sensitivity analysis for US counties

Standard IV Match			Strong IV Match		
β	95% CI		β	95% CI	
0.40	0.20	0.58	0.45	0.21	0.73

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 2002a).

4.2. *Application to the 2000 US election.* Table 5 contains the point estimates and 95% confidence intervals for two of the matches. The first match did not strengthen the instrument and the second match is the based on the stronger instrument that maximized the weak instrument test. For the match where we did not strengthen the instrument, 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. This estimate is also nearly identical to the estimates in the original analysis based on many years of data (Gomez, Hansford and Krause 2007). For the matches which resulted in a stronger instrument, we find that the point estimate is slightly larger in magnitude, at 0.45, while the confidence intervals are also wider, which is expected given the much smaller sample size.

One might reasonably consider whether the estimates in Table 5 are similar to estimates from two-stage least squares applied to all counties, with excess rainfall as an instrument for turnout. Here, we applied two-stage least squares to the unmatched data. We find that two-stage least squares produces a point estimate of 1.86 and a 95% CI [0.57, 3.07] if we do not include covariates, and a point estimate of 3.13 and a 95% CI [2.03, 4.24] with covariates included. Thus two-stage least squares yields much larger estimates. This is not entirely surprising. The two-stage least square estimate is equivalent to

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

which is the well-known Wald (1940) estimator. When the instrument is weak and provides little encouragement, the denominator may be very small, resulting in inflated estimates. The estimates in Table 5 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 rainfall within matched pairs.

4.3. *Sensitivity analysis for generic hidden bias.* In the preceding analysis, we assumed that assignment to encouragement (the instrument) within pairs is as-if random conditional on observed covariates after matching. We first formalize this assumption. Let π_j denote the probability of begin assigned to a value of the instrument for unit j. For two subjects, k and j matched so that observed covariates are similar, $\mathbf{x}_{ik} = \mathbf{x}_{ij}$, we assume that $\pi_j = \pi_k$. However, subjects may differ in the probability of treatment because they differ in terms of some unobserved covariate. That is, it may be the case that we failed to match on an important unobserved binary covariate u such that $\mathbf{x}_{ik} = \mathbf{x}_{ik}$, but possibly $u_{ik} \neq u_{ij}$. If true, the probability of being exposed to treatment may not be constant within matched pairs.

Rosenbaum (2002a, sec. 4.2) proves that we may characterize this probability with a logit model linking the probability of assignment 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 assignment as a function of u_j. For two units, ik and ij with $\mathbf{x}_{ik} = \mathbf{x}_{ij}$, we characterize how they may differ in their odds of assignment with the model above rewritten as: $\pi_{ij}(1-\pi_{ik})/\pi_{ik}(1-\pi_{ij}) = \exp\{\gamma(u_{ij} - u_{ik})\}$.

Now we write $\exp(\gamma) = \Gamma$, and if $\Gamma = 1$ for two matched units, then the units do not differ in their odds of assignment as a function of the unobserved u. For Γ values greater than one, we can place bounds on quantities of interest such as a p-values or point estimates. 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 (Rosenbaum 2002a). Larger values of Γ indicate greater resistance to bias from hidden confounders. For a discussion of different approaches to sensitivity analysis in observational studies see Cornfield et al. (1959), Brumback et al. (2004), Lin, Psaty and Kronmal (1998), Liu, Kuramoto and Stuart (2013), McCandless, Gustafson and Levy (2007), Robins, Rotnitzky and Scharfstein (1999), Rosenbaum (2007), Rosenbaum (2002a), Small (2007), and Small and Rosenbaum (2008)

Baiocchi et al. (2010) find that strengthening the instrument yields an estimate that is more resistant to hidden bias, which is consistent with what statistical theory predicts (Small and Rosenbaum 2008). 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.

Above we tested $H_0 : \beta = 0$ using Wilcoxon's signed rank statistic, U. The sign rank statistic is the sum of S independent random variables where the sth variable equals the sign ±1 with probability 1/2. Now define U^+ to be the sums of S independent random variables where the sth variable takes the sign ±1 with probability p^+ and U^- to be the sums of S independent random variables where the sth variable takes the sign ±1 with probability p^-. Where we define $p^+ = \Gamma/1+\Gamma$ and $p^- = 1/1+\Gamma$. Using these values, we can construct values of U^+ and U^- which form the upper and lower bounds on U for a given value of Γ (Rosenbaum 2002a, sec. 4.3.3). Using U^+ and U^-, we can calculate a set of bounding p-values.

We apply the sensitivity analysis to both the match where no penalties were applied to the instrument and the match where we selected the instrument strength via weak instrument tests. For the weak instrument, we find that the p-value exceeds the conventional 0.05 significance level when $\Gamma = 1.18$. For the stronger instrument, we find that the p-value exceeds the conventional 0.05 significance level when $\Gamma = 1.24$. Therefore, despite the much smaller sample size, we increase resistance to hidden bias in the strong instrument match, which is consistent with statistical theory (Small and Rosenbaum 2008). However, for both matches, a fairly modest amount of hidden bias could explain the results we observe. As such, our conclusions are sensitive to a possible hidden confounder.

5. Discussion. Instruments in natural experiments can often be characterized as weak, in that the instrument provides little encouragement to take the treatment. Baiocchi et al. (2010) developed a matching algorithm produces a set of matched pairs for which the instrument is stronger. Thus analysts need not accept the use of a weak instrument. In applications, however, it can be difficult to know whether the match has produced a sufficiently strong instrument. Using weak instrument tests from the econometric literature, one can map the region of matches where the average within pair distance on the instrument passes a weak instrument test. In our application, we simply chose the combination of penalties and sinks that produced the strongest instrument. We find that rainfall does appear to dissuade voters on election day and this loss of voters tends to help Republican candidates, however, this effect could be easily explained by confounding from an unobserved covariate. While inferences changed little when we strengthened the instrument, conclusions were less sensitive to hidden bias.

References.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* **91** 444–455.

Angrist, J. D. and Pischke, J̈.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics. *Journal of Economic Perspectives* **24** 3-30.

Baiocchi, M., Small, D. S., Lorch, S. and Rosenbaum, P. R. (2010). Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants. *Journal of the American Statistical Association* **105** 1285-1296.

Baiocchi, M., Small, D. S., Yang, L., Polsky, D. and Groeneveld, P. W. (2012). Near/far matching: a study design approach to instrumental variables. *Health Services and Outcomes Research Methodology* **12** 237–253.

Bound, J., Jaeger, D. A. and Baker, R. M. (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** 443-450.

Brumback, B. A., Hernán, M. A., Haneuse, S. J. and Robins, J. M. (2004). Sensitivity Analyses for Unmeasured Confounding Assuming a Marginal Structural Model for Repeated Measures. *Statistics in Medicine* **23** 749–767.

Chamberlain, G. and Imbens, G. (2004). Random effects estimators with many instrumental variables. *Econometrica* **72** 295–306.

Chao, J. C. and Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica* **73** 1673–1692.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. and Wynder, E. (1959). Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions. *Journal of National Cancer Institute* **22** 173-203.

Dunning, T. (2012). *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge University Press, Cambridge, UK.

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

Hansford, T. G. and Gomez, B. T. (2010). Estimating the Electoral Effects of Voter Turnout. *American Political Science Review* **104** 268-288.

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

Holland, P. W. (1988). Causal Inference, Path Analysis, and Recursive Structural Equation Models. *Sociological Methodology* **18** 449-484.

Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica* **62** 467–476.

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.

Keele, L., Titiunik, R. and Zubizarreta, J. (2014). Enhancing a Geographic Regression Discontinuity Design Through Matching to Estimate the Effect of Ballot Initiatives on Voter Turnout. *Journal of the Royal Statistical Society: Series A* **178** 223–239.

Lin, D., Psaty, B. M. and Kronmal, R. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54** 948–963.

Liu, W., Kuramoto, S. J. and Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science* **14** 570–580.

Lorch, S. A., Baiocchi, M., Ahlberg, C. E. and Small, D. S. (2012). The differential impact of delivery hospital on the outcomes of premature infants. *Pediatrics* **130** 270–278.

Lu, B., Zutto, E., Hornik, R. and Rosenbaum, P. R. (2001). Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse. *Journal of the American Statistical Association* **96** 1245–1253.

McCandless, L. C., Gustafson, P. and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in medicine* **26** 2331–2347.

Nagler, J. (1991). The Effect of Registration Laws and Education on United-States Voter Turnout. *American Political Science Review* **85** 1393-1405.

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).

Robins, J. M., Rotnitzky, A. and Scharfstein, D. (1999). Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In *Statistical Models in Epidemiology: The Environment and Clinical Trials.* (E. Halloran and D. Berry, eds.) 1–92. Springer.

Rosenbaum, P. R. (1984). The Consequences of Adjusting For a Concomitant Variable That Has Been Affected By The Treatment. *Journal of The Royal Statistical Society Series A* **147** 656–666.

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

Rosenbaum, P. R. (1999). Using quantile averages in matched observational studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **48** 63–78.

Rosenbaum, P. R. (2002a). *Observational Studies*, 2nd ed. Springer, New York, NY.

Rosenbaum, P. R. (2002b). Covariance Adjustment In Randomized Experiments and Observational Studies. *Statistical Science* **17** 286-387.

Rosenbaum, P. R. (2005). Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Obervational Studies. *The American Statistician* **59** 147-152.

Rosenbaum, P. R. (2007). Sensitivity Analysis for m-Estimates, Tests, and Confidence Intervals in Matched Observational Studies. *Biometrics* **63** 456–464.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer-Verlag, New York.

Rosenzweig, M. R. and Wolpin, K. I. (2000). Natural 'Natural Experiments' in Economics. *Journal of Economic Literature* **38** 827-74.

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

Rubin, D. B. (1986). Which Ifs Have Causal Answers. *Journal of the American Statistical Association* **81** 961-962.

Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association* **102** 1049–1058.

Small, D. and Rosenbaum, P. R. (2008). War and Wages: The Strength of Instrumental Variables and Their Sensitivity to Unobserved Biases. *Journal of the American Statistical Association* **103** 924-933.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association* **101** 1398–1407.

Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak

Instruments. *Econometrica* **65** 557–586.

Stock, J. H. and Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. In *Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg* (D. W. K. Andrews and J. H. Stock, eds.) 5 Cambridge University Press.

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

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

Zigler, C. M., Dominici, F. and Wang, Y. (2012). Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics* **13** 289-302.

Zubizarreta, J. R., Small, D. S. and Rosenbaum, P. R. (2014). Isolation In the Construction of Natural Experiments. *The Annals of Applied Statistics* **Forthcoming.**

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S. and Rosenbaum, P. R. (2013). Stronger Instruments via Interger Programming in an Observational Study of Late Preterm Birth Outcomes. *Annals of Applied Statistics* **7** 25–50.

Department of Political Science,
211 Pond Lab, Penn State University,
University Park, PA 16802
E-mail: ljk20@psu.edu

Department of Political Science,
2140 Derby Hall, Ohio State University,
Columbus, OH 43210
E-mail: morgan.746@osu.edu