Opiates for the Matches:

Matching Methods for Causal Inference*

Jasjeet S. Sekhon†

Annual Review of Political Science, 2009

*I thank Jake Bowers, David Freedman, Ben Hansen, Shigeo Hirano, Walter Mebane, Jr., Donald Rubin, Jonathan Wand, and Rocío Titiunik for valuable comments and advice. I also thank an anonymous reviewer for extensive and extremely helpful comments. All errors are my responsibility.

†Associate Professor, Travers Department of Political Science, sekhon@berkeley.edu, http://sekhon.berkeley.edu/, Survey Research Center, 2538 Channing Way, UC Berkeley, Berkeley, CA, 94720.

Abstract

In recent years there has been a burst of innovative work on methods for estimating causal effects using observational data. Much of this work has extended and brought a renewed focus on old approaches such as matching, which is the focus of this review. The new developments highlight an old tension in the social sciences: a focus on research design versus a focus on quantitative models. This realization along with the renewed interest in field experiments has marked the return of foundational questions as opposed to a fascination with the latest estimator. I use studies of get-out-the-vote interventions to exemplify this development. Without an experiment, natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive.

Key words: Causal Inference, Matching, Neyman-Rubin Model

1 Introduction

Although the quantitative turn in the search for causal inferences is over a century old in the social sciences, in recent years there has been a renewed interest in the problems associated with making causal inferences using such methods. These recent developments highlight the tensions in the quantitative tradition that have been present from the beginning. There are a number of conflicting approaches, which overlap but have important distinctions. I focus here on three of them: the experimental, the model based, and the design based.

The first, which is surprisingly old, is the use of randomized experiments, which in political science may go back to Gosnell (1927).[1] Whether Gosnell randomized or not, Eldersveld (1956) certainly did when he conducted a randomized field experiment to study the effectiveness of mail, telephone and house-to-house canvassing on voter mobilization. But even with randomization, there is ample disagreement and confusion about exactly how such data should be analyzed—for example, "is adjustment by multivariate regression unbiased?". There are also concerns about external validity and whether experiments can be used to answer "interesting" or "important" questions. This latter concern appears to be common among social scientists and is sometimes harshly put. As one early and suspicious reviewer of experimental methods in the social sciences noted: *parturiunt montes, nascetur ridiculus mus* (Mueller 1945).[2] For observational data analysis, however, the disagreements are sharper.

By far the dominant method of making causal inferences in the quantitative social sciences is model based, and the most popular model is multivariate regression. This tradition is also surpris-

[1] Gosnell many not have actually used randomization (Green and Gerber 2002). Part of the issue is that Gosnell's 1924 Get-Out-The-Vote experiment, described in his 1927 book, was conducted one year before Fisher's 1925 book and eleven years before Fisher's famous 1935 book on experimental design. Therefore, unsurprisingly, Gosnell's terminology is non-standard and leads to some uncertainty about exactly what was done. Wand (2008), in a note reviewing the question, hypothesizes that of the 1924 experiment, "the description of the design and his [Gosnell's] arguments against possible confounders imply that Gosnell used a cluster design which *randomly* assigned one half of the residents within each deterministically chosen district to either treatment or control" (emphasis in original). Wand notes that the description in Gosnell (1948) is clearer, but it may not describe the original study. A definitive answer to the question requires a close examination of Gosnell's papers at the University of Chicago.

[2] "The mountains are in labor, a ridiculous mouse will be brought forth" which is a quotation from Horace, Epistles, Book II, Ars Poetica (The Art of Poetry). Horace is observing that some poets make great promises which result in little.

ingly old; the first use of regression to estimate treatment effects (as opposed to simply fitting a line through data) was Yule's (1899) investigation into the causes of changes in pauperism in England.[3]

The third tradition focuses on design. Examples abound, but they can be broadly categorized as natural experiments or Regression-Discontinuity (RD) designs. They share in common an assumption that found data, not part of an actual field experiment, has some "as if" random component: that the assignment to treatment can be treat "as if" it were randomly assigned or can be so treated after some covariate adjustment. From the beginning, some natural experiments were analyzed as if they were actual experiments (e.g., difference of means), others by matching methods (e.g., Chapin 1938) and yet others (many, many others) by instrumental variables (e.g., Wright 1928).[4] A central criticism of natural experiments is that they are not randomized experiments. In most cases, the "as if" random assumption is implausible. For reviews see Dunning (2008) and Rosenzweig and Wolpin (2000).

Regression-discontinuity was first proposed by Thistlethwaite and Campbell (1960). They proposed RD as an alternative to what they called "ex post facto experiments," or what we today would call natural experiments analyzed by *matching* methods. More specifically, they proposed RD as an alternative to matching methods and other "as if" (conditionally) random experiments outlined by Chapin (1938) and Greenwood (1945) where the assignment mechanism is not well understood. In the case of RD, the researcher finds a sharp breakpoint that makes seemingly random distinctions between units that receive treatment and those that do not.

Where does matching fit in? As we shall see, it depends on how it is used.

One of the innovative intellectual developments over the past few years has been to unify all of these methods into a common mathematical and conceptual language, that of the Neyman-Rubin model (Neyman 1923 [1990]; Rubin 1974). Although randomized experiments and matching estimators have long been tied to the model, recently instrumental variables (Angrist, Imbens, and Rubin 1996) and regression discontinuity (Lee 2008) have also been so tied. This leads to an in-

[3]By Yule (1899), the understanding of regression had evolved from what Stigler (1990) calls the Gauss-Laplace synthesis. For example, Yule understood the regression coefficient in terms of variances and covariances.

[4]For an interesting note on who invented instrumental variable regression see Stock and Trebbi (2003).

teresting unity of thought that makes clear that the Neyman-Rubin model is the core of the causal enterprise, and that the various methods and estimators consistent with it, although practically important, are of secondary interest. These are fighting words because all of these techniques, particularly the clearly algorithmic ones such as matching, can be used without any ties to the Neyman-Rubin model or causality. In such cases matching becomes nothing more than a non-parametric estimator, a method to be considered along side CART (Breiman, Friedman, Stone, and Olshen 1984), BART (Chipman, George, and McCulloch 2006), kernel estimation, and a host of others. Matching becomes simply a way to lessen model dependence, not a method for estimating causal effects per se. For causal inference, issues of design are of utmost importance; a lot more is needed than just an algorithm. Like other methods, matching algorithms can always be used, and they usually are, even when design issues are ignored in order to obtain a non-parametric estimate from the data. Of course, in such cases what exactly has been estimated is unclear.

The Neyman-Rubin model has radical implications for work in the social sciences given current practices. According to this framework, much of the quantitative work which is done that claims to be causal, is not well posed. The questions asked are too vague, and the design is hopelessly compromised by, for example, conditioning on post-treatment variables (Cox 1958, §4.2, Rosenbaum 2002, 73–74).

The radical import of the Neyman-Rubin model may be highlighted by using it to determine how regression estimators behave when fitted to data from randomized experiments. Randomization does not justify the regression assumptions (Freedman 2008b,c). Without additional assumptions, multiple regression is not unbiased. The variance estimates from multiple regression may be arbitrarily too large or too small, even asymptotically. And for logistic regression, matters only become worse (Freedman 2008d). These are fearful conclusions. These pathologies occur even with randomization. That's supposed to be the easy case.

Although the Neyman-Rubin model is currently the most prominent, and I focus on it in this review, there have obviously been many other attempts to understand causal inference. For a review see Brady (2008). In recent years, an alternative which is growing in prominence is Pearl's

(2000) work on non-parametric structural equations models. For a critique see Freedman (2004). Pearl's approach is a modern reincarnation of an old enterprise which has a rich history including foundational work on causality in systems of structural equations by the political scientist Herbert Simon (1953). Haavelmo (1943) was the first to precisely examine issues of causality in the context of linear structural equations with random errors.

On matching itself, there is no consensus on how exactly matching ought to be done, how to measure the success of the matching procedure, and whether or not matching estimators are sufficiently robust to misspecification so as to be useful in practice (Heckman, Ichimura, Smith, and Todd 1998). To illuminate issues of general interest, I review a prominent exchange in the political science literature involving a set of Get-Out-The-Vote (GOTV) field experiments and the use of matching estimators (Arceneaux, Gerber, and Green 2006; Gerber and Green 2000, 2005; Imai 2005; Hansen and Bowers In Press).

The matching literature is growing rapidly, so it is impossible to summarize it in a brief review. I focus on design issues and less on the technical details of exactly how matching should be done, although the basics are reviewed. For an excellent review of recent developments in methods for program evaluation see Imbens and Wooldridge (2008). For additional reviews of the matching literature see Morgan and Harding (2006), Morgan and Winship (2007), Rosenbaum (2005), and Rubin (2006).

This review is organized as follows. In Section 2, I describe the Neyman-Rubin model, its intellectual history, and various implications of the model for research practices. Examining the contested intellectual history of the Neyman-Rubin model helps to highlight its core principles. In Section 3, I briefly outline some common matching methods. In Section 4, I review the GOTV matching controversy. Section 5 concludes.

2 Neyman-Rubin Causal Model

The Neyman-Rubin framework has become increasingly popular in many fields including statistics (Holland 1986; Rubin 2006, 1974; Rosenbaum 2002), medicine (Christakis and Iwashyna 2003; Rubin 1997), economics (Abadie and Imbens 2006a; Galiani, Gertler, and Schargrodsky 2005; Dehejia and Wahba 2002, 1999), political science (Bowers and Hansen 2005; Imai 2005; Sekhon 2004), sociology (Morgan and Harding 2006; Diprete and Engelhardt 2004; Winship and Morgan 1999; Smith 1997) and even law (Rubin 2001). The framework originated with Neyman's (1923 [1990]) model which is non-parametric for a finite number of treatments where each unit has two potential outcomes for each treatment, one if the unit is treated and the other if untreated. A causal effect is defined as the difference between the two potential outcomes, but only one of the two potential outcomes is observed. Rubin (1974, 2006) developed the model into a general framework for causal inference with implications for observational research. Holland (1986) wrote an influential review article that highlighted some of the philosophical implications of the framework. Consequently, instead of the "Neyman-Rubin model," the model is often simply called the Rubin causal model (e.g., Holland 1986) or sometimes the Neyman-Rubin-Holland model (e.g., Brady 2008) or the Neyman-Holland-Rubin model (e.g., Freedman 2006).

The intellectual history of the Neyman-Rubin model is the subject of some controversy (e.g., Freedman 2006; Rubin 1990; Speed 1990). Neyman's 1923 article never mentions the random assignment of treatments. Instead, the original motivation was an urn model, and the explicit suggestion to use the urn model to physically assign treatments is absent from the paper (Speed 1990).[5] It was left to R. A. Fisher in the 1920s and 1930s to note the importance of the physical act of randomization in experiments. Fisher first did this in the context of experimental design in his 1925 book, expanded on the issue in a 1926 article for agricultural researchers, and developed it more fully and for a broader audience in his 1935 book *The Design of Experiments*.[6] As Reid (1982, 45) notes of Neyman:

[5] An urn model is based on an idealized thought experiment in which colored balls in an urn are drawn randomly. Using the model does not imply that treatment should be physically assigned in a random fashion.

[6] For more on Fisher's role in the advocacy of randomization see Armitage (2003); Hall (2007); Preece (1990).

> On one occasion, when someone perceived him [Neyman] as anticipating the English statistician R. A. Fisher in the use of randomization, he objected strenuously:
> "I treated *theoretically* an unrestrictedly randomized agricultural experiment and the randomization was considered as a prerequisite to probabilistic treatment of the results. This is not the same as the recognition that without randomization an experiment has little value irrespective of the subsequent treatment. The latter point is due to Fisher, and I consider it as one of the most valuable of Fisher's achievements."[7]

This gap between Neyman and Fisher points to the fact that there was something absent from the Neyman mathematical formulation in 1923, which was added later, even though the symbolic formulation was complete in 1923. What those symbols *meant* changed. And in these changes lies what is causal about the Neyman-Rubin model—i.e., a focus on the mechanism by which treatment is assigned.

The Neyman-Rubin model is more than just the math of the original Neyman model. Obviously, it does not rely upon an urn model motivation for the observed potential outcomes, but, for experiments, a motivation based on the random assignment of treatment. And for observational studies, one relies on the assumption that the assignment of treatment can be treated as-if it were random. In either case, the mechanism by which treatment is assigned is of central importance. And the realization that the primacy of the assignment mechanism holds true for observational data no less than for experimental, is due to Rubin (1974). This insight has been turned into a motto: "no causation without manipulation" (Holland 1986).

Although the original article was written in Polish, Neyman's work was known in the English speaking world (Reid 1982), and in 1938 Neyman moved from Poland to Berkeley. It is thus unsurprising that the Neyman model quickly became the standard way of describing potential outcomes of randomized experiments: for example, Pitman (1937); Welch (1937); McCarthy (1939); Anscombe (1948); Kempthorne (1952, 1955). The most complete discussion I know of before Rubin's work is Scheffé (1956). And a simplified version of the model even appears in Hodges's

[7] Also see Rubin (1990, 477).

and Lehmann's introductory textbook in 1964 (section 9.4).[8]

The basic setup of the Neyman model is very simple. Let Y_{i1} denote the potential outcome for unit i if the unit receives treatment, and let Y_{i0} denote the potential outcome for unit i in the control regime. The treatment effect for observation i is defined by $\tau_i = Y_{i1} - Y_{i0}$. Causal inference is a missing data problem because Y_{i1} and Y_{i0} are never both observed. This remains true regardless of the methodology used to make inferential progress—regardless of whether we use quantitative or qualitative methods of inference. The fact remains that we cannot observe both potential outcomes at the same time.

Some set of assumptions have to be made to make progress. The most compelling are offered by a randomized experiment. Let T_i be a treatment indicator: 1 when i is in the treatment regime and 0 otherwise. The observed outcome for observation i is then:

$$Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}. \tag{1}$$

Note that in contrast to the usual regression assumptions, the potential outcomes, Y_{i0} and Y_{i1}, are fixed quantities and not random variables, and that Y_i is only random because of treatment assignment.[9]

2.1 Experimental Data

In principle, if assignment to treatment is randomized, causal inference is straightforward because the two groups are drawn from the same population by construction, and treatment assignment is independent of all baseline variables. The distributions of both observed and unobserved variables between treatment and control groups are equal—i.e., the distributions are *balanced*. This occurs with arbitrarily high probability as the sample size grows large.

[8]The philosopher David Lewis (1973) is often cited for hypothetical counterfactuals and causality, and it is sometimes noted that he predated, by a year, Rubin (1974). The Neyman model predates Lewis.

[9]Extensions to the case of multiple discrete treatment are straightforward (e.g., Imbens 2000, Rosenbaum 2002 300–302). Extensions to the continuous case are possible but lose the nonparametric nature of the Neyman model, see Imai and van Dyk (2004).

Treatment assignment is independent of Y_0 and Y_1—i.e., $\{Y_{i0}, Y_{i1} \perp\!\!\!\perp T_i\}$, where $\perp\!\!\!\perp$ denotes independence. In other words, the distributions of both of the potential outcomes (Y_0, Y_1) are the same for treated ($T = 1$) and control ($T = 0$). Hence, for $j = 0, 1$

$$E(Y_{ij} \mid T_i = 1) = E(Y_{ij} \mid T_i = 0), \tag{2}$$

where the expectation is taken over the distribution of treatment assignments. This equation states that the distributions of potential outcomes in treatment and control are the same in expectation. But for treatment observations one observes T_{i1} and for control observations T_{i0}. Treatment status filters which of the two potential outcomes we observe (Equation 1), but does not change them.

The average treatment effect (ATE) is defined to be:

$$\begin{aligned} \tau &= E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 0) \\ &= E(Y_i \mid T_i = 1) - E(Y_i \mid T_i = 0) \end{aligned} \tag{3}$$

Equation 3 can be estimated consistently by simply taking the difference between two sample means because randomization ensures that the potential outcomes in treatment and control groups have the same distributions in expectation. This implies that randomization ensures that assignment to treatment will not be associated with any potentially confounding variable—i.e., with any pre-treatment variable associated with the outcome.

One of the assumptions which randomization by itself does not justify is that "the observation on one unit should be unaffected by the particular assignment of treatments to the other units" (Cox 1958, §2.4). This "no interference between units" is often called the Stable Unit Treatment Value Assumption (SUTVA). SUTVA implies that the potential outcomes for a given unit do not vary with the treatments assigned to any other unit, and that there are not different versions of treatment (Rubin 1978). SUTVA is a complicated assumption which is all too often ignored.

Brady (2008) describes a randomized welfare experiment in California where SUTVA is violated. In the experiment, teenage girls in the treatment group had their welfare checks reduced if

they failed to obtain passing grades in school. Girls in the control group did not face the risk of reduced payments. However, some girls in the control group thought that they were in the treatment group probably because they knew girls in treatment (Mauldon, Malvin, Stiles, Nicosia, and Seto 2000). Therefore, the experiment probably underestimated the effect of the treatment.

Some researchers erroneously think SUTVA is another term for the assumption usually made in regression models that the disturbances of different observations are independent of one another. A hint of the problem can be seen by noting that OLS is still unbiased under the usual assumptions even if multiple draws from the disturbance are not independent of each other. When SUTVA is violated, however, an experiment will not generally yield unbiased estimates (Cox 1958). In the usual regression setup, the correct specification assumption deals with SUTVA violations: it is implicitly assumed that if there are SUTVA violations, we have the correct model for them so that conditional independence holds—i.e., $E(\epsilon \mid X) = 0$, where ϵ is the regression disturbance and X are the observed variables.

Even with randomization, the usual OLS regression assumptions are not satisfied. Indeed, without further assumptions, the multiple regression estimator is biased. Asymptotically the bias vanishes in some cases, but need not with cluster randomized experiments (Middleton 2008). The regression standard errors can be severely biased, and the multiple regression estimator may have higher asymptotic variance than simply estimating Equation 3. For details see Freedman (2008b,c). Intuitively, the problem is that generally, even with randomization, the treatment indicator and the disturbance will be strongly correlated. Randomization does not imply, as OLS assumes, a linear additive treatment effect where the coefficients are constant across units. Random effects do not solve the problem. Linear additivity remains, and the heterogeneity of the causal effect must be modeled. But the model may be wrong. For example, the effect may not vary normally as is commonly assumed, and it may be strongly related to other variables in the model. Researchers should be extremely cautious about using multiple regression to adjust experimental data. Unfortunately, there is a tendency to use it freely. One supposes that this is yet another sign, as if one more were needed, of how ingrained the regression model is in our quantitative practice.

Unlike multiple regression, random assignment of treatment is sufficient for simple bivariate regression to be an unbiased estimator for Equation 3. The simple regression estimator is obtained by running a regression of the observed response Y on the assignment variable T with an intercept. The standard errors of this estimator are, however, generally incorrect because the standard regression formulas assume homoscedasticity. Alternative variance estimators that adjust for heteroscedasticity may be used.[10]

The only thing stochastic in the Neyman-Rubin framework is the assignment to treatment. The potential outcomes are fixed. This is exactly the opposite of many econometric treatments where all of the regressors (including the treatment indicator) are considered to be fixed, and the response variable Y is considered to be a random variable with a given distribution. None of that is implied by randomization and indeed randomization explicitly contradicts it because one of the regressors (the treatment indicator) is explicitly random. Adding to the confusion is the tendency of some texts to refer to the fixed regressors design as an experiment when that cannot possibly be the case.

In many modern treatments of OLS, X is stochastic, but that raises additional questions. What makes the X covariates (other than the randomly assigned treatment indicator) random? And if the data are a random sample (so, clearly, X is random), then there are two distinct sources of randomness: (i) treatment assignment; (ii) sampling from a population. These are distinct entities and one could be interested in either sample or population estimates—e.g., Sample Average Treatment Effects (SATE) or Population Average Treatment Effects (PATE). Sample estimates ignore the second source of randomness, and the population estimates take both into account. In the case of random sampling, SATE generally has less variance than PATE but certainly no more (Imbens 2004). Without assumptions in addition to random assignment and random sampling, one is not led to the usual regression variance formulas.

A parallel argument holds if one wants to consider the potential outcomes to be random and not fixed. What is the source and model of this randomness? Without additional information, it is

[10] An obvious alternative is to use the following variance estimator: $\frac{\hat{v}_t}{n_t} + \frac{\hat{v}_c}{n_c}$, where $\hat{v}_t$ is the sample variance for the treatment observations, n_t is the number of treatment observations, and the subscript c denotes analogous quantities for the control group.

most natural to consider that the potential outcomes are fixed because in a randomized experiment all that we *know* is random is treatment assignment. In the case of random potential outcomes, one can always conduct an analysis conditional on the data at hand, like SATE above, which ignores the second source of randomness. Of course, the conditional inference (e.g., SATE) may lead to a different inference than the unconditional inference. Without assumptions (such as random sampling), the sample contains no information about the PATE beyond the SATE. Note that if the potential outcomes are random, but we condition on the observed potential outcomes and so treat them as fixed, questions about the role of conditioning and inference arise, which go back to Neyman and Fisher. If the random error is independent of treatment assignment, this situation is analogous to the case of a 2×2 table where one margin is fixed and we analyze the data as if both margins are fixed (Lehmann 1993, Rosenbaum 2005, §2.5–2.9).

Even in an experimental setup, much can go wrong that requires statistical adjustment (e.g., Barnard, Frangakis, Hill, and Rubin 2003). A common problem is compliance. For example, a person assigned to treatment may refuse it. This person is said to have crossed over from treatment to control. And a person assigned to control may find some way to receive treatment nevertheless, which is another form of crossover.

When there are compliance issues, Equation 3 defines the Intention-To-Treat (ITT) estimand. Although the concept of ITT dates earlier, the phrase probably first appeared in print in Hill (1961, 259). Moving beyond the ITT to estimate the effect of treatment on the units which actually received treatment can be difficult. ITT measures the effect of assignment rather than treatment itself. And estimates of ITT are unbiased even with crossover. The obvious benefit is that ITT avoids bias by taking advantage of the experimental design.

The simplest compliance problem is where every unit assigned to control accepts control, but where some units assigned to treatment decline treatment, and follow the control protocol instead. This is called single crossover. In this case, the Neyman-Rubin model can easily handle the issue. Progress is made by assuming that there are two types of units: compliers and never-treat. A complier follows her assignment to either treatment or control. Compliers have two potential

outcomes which are observed as in Equation 1. However, a never-treat unit is assumed to have only one response, and this response is observed regardless if the unit is randomized to receive treatment or control.

With this simple model in place, we have five different parameters:

(i) the proportion of compliers in the experimental population (α)

(ii) the average response of compliers assigned to treatment ($\bar{W}$)

(iii) the average response of compliers assigned to control ($\bar{C}$)

(iv) the difference between $\bar{W}$ and $\bar{C}$, which is the average effect of treatment on the compliers ($\bar{R}$)

(v) the average response of never-treat units assigned to control ($\bar{Z}$)

All five of these parameters can be estimated. α can be estimated by calculating the proportion of compliers observed in the treatment group. Because of randomization, this proportion is an unbiased estimate of the proportion of compliers in control as well. The average response of compliers to treatment, $\bar{W}$, is simply the average response of compliers in the treatment group. And $\bar{Z}$, the average response of never-treat units to control, is estimated by the average response among units in the treatment group who refused treatment.

This leaves $\bar{C}$ and $\bar{R}$. For $\bar{R}$, note that the control group contains a mix of compliers and never-treat units. We do not know the type of any given unit in control, but we know (in expectation) the *proportion* of each there must be in control because we can estimate this proportion in the treated group.

Recall that α denotes the proportion of compliers in the experimental population, and assume that $\alpha > 0$. Under the model, the proportion of never-treat units must be $1 - \alpha$. Denote the average observed responses in treatment and control by $\bar{Y}^t$, $\bar{Y}^c$, these are sample quantities which are directly observed. Since the treatment and control groups are exchangeable because of random

assignment,

$$E(\bar{Y}^c) \;=\; \alpha\bar{C} + (1-\alpha)\bar{Z}.$$

Therefore,

$$\bar{C} \;=\; \frac{E(\bar{Y}^c) - (1-\alpha)\bar{Z}}{\alpha}.$$

An obvious estimator for $\bar{C}$ is

$$\hat{\bar{C}} \;=\; \frac{\bar{Y}^c - (1-\hat{\alpha})\hat{\bar{Z}}}{\hat{\alpha}}.$$

Then the only remaining quantity is $\bar{R}$, the average effect of treatment on the compliers—i.e., the Effect of Treatment on the Treated (ETT). This can be estimated by

$$\hat{\bar{W}} - \hat{\bar{C}} \;=\; \frac{\bar{Y}^t - \bar{Y}^c}{\hat{\alpha}}. \tag{4}$$

Note how simple and intuitive Equation 4 is. The estimated average effect of treatment on the treated is calculated by dividing the ITT estimator by the compliance rate. Because this rate is less than or equal to 1 and by assumption above 0, ETT will be greater than or equal to ITT, and both will have the same sign.

Equation 4 is the same as two-stage least squares (2SLS) where the instrument is the random assignment to treatment. The canonical citation for this estimator is Angrist et al. (1996); they provide a more general derivation.[11] For other discussions see Angrist and Imbens (1994), Bloom (1984), Freedman (2006), and Sommer and Zeger (1991).

When the compliance problem has a more complicated structure (e.g., when there is two-way crossover), it is difficult to make progress without making strong structural assumptions (Freedman 2006). We return to the issue of compliance when we discuss the get-out-the-vote controversy in Section 4.

[11] Note that the discussion above implicitly satisfies the assumptions outlined in Angrist et al. (1996).

2.2 Observational Data

In an observational setting, unless something special is done, treatment and non-treatment groups are almost never balanced because the two groups are not ordinarily drawn from the same population. Thus, a common quantity of interest is the average treatment effect for the treated (ATT):

$$\tau \mid (T = 1) \quad = \quad E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 1), \tag{5}$$

where the expectation is taken over the distribution of treatment assignments. Equation 5 cannot be directly estimated because Y_{i0} is not observed for the treated. Progress can be made by assuming that selection for treatment depends on observable covariates X. Then, one can assume that conditional on X, treatment assignment is unconfounded. In other words, the conditional distributions of the potential outcomes are the same for treated and control: $\{Y_0, Y_1 \perp\!\!\!\perp T\} \mid X$.

Following Rosenbaum and Rubin (1983), we say that treatment assignment is strongly ignorable given a vector of covariates X if unconfoundedness and common overlap hold:

$$\begin{aligned} &\{Y_0, Y_1 \perp\!\!\!\perp T\} \mid X \\ 0 < &Pr(T = 1 \mid X) < 1 \end{aligned}$$

for all X. Heckman et al. (1998) shows that for ATT, the unconfoundedness assumption can be weakened to conditional mean independence between the potential outcomes Y_{ij} and the treatment indicator T_i given X_i.[12]

The common overlap assumption ensures that some observed value of X does not deterministically result in a given observation being assigned to treatment or control. If such deterministic treatment assignments were to occur, it would not be possible to identify the treatment effect. For example, if women were never treated and men always treated, it would not be possible to obtain

[12] Also see Abadie and Imbens (2006a).

an unbiased estimate of the average treatment effect without an additional assumption.[13]

Given strong ignorability, following Rubin (1974, 1977) we obtain

$$E(Y_{ij} \mid X_i, T_i = 1) = E(Y_{ij} \mid X_i, T_i = 0). \tag{6}$$

Equation 6 is the observational equivalent of Equation 2. Equation 6 is a formalization of the as-if random assumption made in observational studies. Once some observable variables have been conditioned upon, analysis can continue as-if treatment were randomly assigned. A key goal is to obtain results for observational data that were demonstrated to hold given random assignment in the previous section.

By conditioning on observed covariates, X_i, treatment and control groups are balanced—i.e., the distributions of the potential outcomes between treatment and control groups are the same. When it comes to potential outcomes, the only difference between the two groups is the potential outcomes we observe, Y_i or Y_0. The average treatment effect for the treated is estimated as

$$\tau \mid (T = 1) \quad = \quad E\left\{E(Y_i \mid X_i, T_i = 1) - E(Y_i \mid X_i, T_i = 0) \mid T_i = 1\right\}, \tag{7}$$

where the outer expectation is taken over the distribution of $X_i \mid (T_i = 1)$, which is the distribution of X in the treated group.

Note that the ATT estimator is changing how individual observations are weighted, and that observations which are outside of common support receive zero weights. That is, if some covariate values are only observed for control observations, those observations will be irrelevant for estimating ATT and are effectively dropped. Therefore, the overlap assumption for ATT only requires that the support of X for the treated observations be a subset of the support of X for control observations. More generally, one would also want to drop treatment observations if they have covariate values which do not overlap with control observations (Crump, Hotz, Imbens, and Mitnik 2006).

[13] We could assume that being a woman or man were independent of the potential outcomes. Women in control could then be valid counterfactuals for men in treatment given the Y of interest. Such additional exclusion assumptions are not required if strong ignorability holds.

In such cases, it is unclear exactly what estimand one is estimating because it is no longer ATT as some treatment observations have been dropped along with some control observations.

It is often jarring for people to observe that observations are being dropped because of a lack of covariate overlap. Our intuition against dropping observations comes from what happens with experimental data, where homogeneity between treatment and control is guaranteed by randomization so a larger sample is obviously better than a smaller one. But with observational data, dropping observations which are outside of common support not only reduces bias but can also reduce the variance of our estimates. This may be counter intuitive, but note that our variance estimates are a function of both sample size and unit heterogeneity—e.g., in the regression case, of the sample variance of X and the mean square error. Dropping observations outside of common support and conditioning as in Equation 7 helps to improve unit homogeneity and may actually reduce our variance estimates (Rosenbaum 2005). Rosenbaum also shows that, with observational data, minimizing unit heterogeneity reduces both sampling variability and sensitivity to unobserved bias. With less unit heterogeneity, larger unobserved biases need to exist to explain away a given effect. And although increasing the sample size reduces sampling variability, it does little to reduce concerns about unobserved bias. Thus, maximizing unit homogeneity to the extent possible is an important task for observational methods.[14]

The key assumption being made here is strong ignorability. Even thinking about this assumption presupposes some rigor in the research design. For example, is it clear what is pre- and what is post- treatment? If not, one is unable to even form the relevant questions. The most useful of which may be the one suggested by Dorn (1953, 680) who proposed that the designer of every observational study should ask "[h]ow would the study be conducted if it were possible to do it by controlled experimentation?" This clear question also appears in Cochran's famous Royal Statistical Society discussion paper on the planning of observational studies of human populations (1965). And Dorn's question has become one which researchers in the tradition of the Neyman-

[14]There is a trade-off between having a fewer number of more homogeneous observations and larger number of more heterogeneous observations. Whether dropping a given observation actually increases the precision of the estimate dependents on exactly how different this observation is from the observations that remain and how sensitive the estimator is to heterogeneity. See Rosenbaum (2005) for formal details.

Rubin model ask themselves and their students. The question forces the researcher to focus on a clear manipulation and then on the selection problem at hand. Only then can one even begin to think clearly about how plausible the strong ignorability assumption may or may not be. It is fair to say that without answering Dorn's question, one is unsure what the researcher wants to estimate. Since most researchers do not propose an answer to this question, it is difficult to think clearly about the underlying assumptions being made in most applications in the social sciences because one is unclear as to what precisely the researcher is trying to estimate.

For the moment let us assume that the researcher has a clear treatment of interest, and a set of confounders which may reasonably ensure conditional independence of treatment assignment. At that point, one needs to condition on these confounders denoted by X. But we must remember that selection on observables is a large concession, which should not be made lightly. It is of far greater relevance than the technical discussion which follows on the best way to condition on covariates.

In other words, the identification assumption for both OLS and matching is the same: selection on observables. Both also rely on SUTVA and have similar restrictions on the use of post-treatment variables. Given that, for all of their differences, they have more in common than most applied researchers in political science realize. The focus then should be more on the identification assumption—e.g., selection on observables—than is often the case in the literature. Authors, even when they have natural experiments, spend insufficient effort justifying this assumption.[15] Obviously, matching is non-parametric while OLS is not. This is an important distinction because asymptotically matching does not make a functional form assumption in addition to the selection of observables assumption (Abadie and Imbens 2006a). OLS, however does make additional assumptions; it assumes linear additivity.

[15]For a review and evaluation of a number of natural experiments and their as-if random assumptions see Dunning (2008).

3 Matching Methods

The most straightforward and nonparametric way to condition on X is to exactly match on the covariates. This is an old approach going back to at least Fechner (1966 [1860]), the father of psychophysics. This approach is often impossible to implement in finite samples if the dimensionality of X is large—i.e., exact matches are not found in a given sample. And exact matching is not possible to implement even asymptotically if X contains continuous covariates. Thus, in general, alternative methods must be used.

Various forms of matching have been used for some time, for example (Chapin 1938; Cochran 1953; Greenwood 1945). Two common approaches today are propensity score matching (Rosenbaum and Rubin 1983) and multivariate matching based on Mahalanobis distance (Cochran and Rubin 1973; Rubin 1979, 1980).

3.1 Mahalanobis and Propensity Score Matching

The most common method of multivariate matching is based on Mahalanobis distance (Cochran and Rubin 1973; Rubin 1979, 1980). The Mahalanobis distance between any two column vectors is:

$$md(X_i, X_j) \quad = \left\{(X_i - X_j)' S^{-1} (X_i - X_j)\right\}^{\frac{1}{2}}$$

where S is the sample covariance matrix of X. To estimate ATT, one matches each treated unit with the M closest control units, as defined by this distance measure, $md()$. Matching with replacement results in the estimator with the lowest conditional bias (Abadie and Imbens 2006a).[16] If X consists of more than one continuous variable, multivariate matching estimates contain a bias term which does not asymptotically go to zero at $\sqrt{n}$ (Abadie and Imbens 2006a).

An alternative way to condition on X is to match on the probability of assignment to treatment,

[16] Alternatively, one can use optimal full matching (Hansen 2004; Rosenbaum 1991) instead which may have lower variance. But this decision is a separate one from the choice of a distance metric.

known as the propensity score.[17] As one's sample size grows large, matching on the propensity score produces balance on the vector of covariates X (Rosenbaum and Rubin 1983).

Given strong ignorability, Rosenbaum and Rubin (1983) prove

$$\tau \mid (T = 1) \quad = \quad E\left\{E(Y_i \mid e(X_i), T_i = 1) - E(Y_i \mid e(X_i), T_i = 0) \mid T_i = 1\right\},$$

where the outer expectation is taken over the distribution of $e(X_i) \mid (T_i = 1)$. Under these assumptions, the propensity score can be used to provide an unbiased estimate of ATE as well.

Propensity score matching usually involves matching each treated unit to the nearest control unit on the unidimensional metric of the propensity score vector.[18] Since the propensity score is generally unknown, it must be estimated. If the propensity score is estimated by logistic regression, as is typically the case, much is to be gained by matching not on the predicted probabilities (bounded between zero and one) but on the linear predictor: $\hat{\mu} = X\hat{\beta}$. Matching on the linear predictor avoids compression of propensity scores near zero and one (Rosenbaum and Rubin 1985). Moreover, the linear predictor is often more nearly normally distributed which is of some importance given the Equal Percent Bias Reduction theoretical results discussed below.

Mahalanobis distance and propensity score matching can be combined in various ways (Rubin 2001). Rosenbaum and Rubin (1985) show that, in finite samples, it is useful to combine the two matching methods because doing so results in a greater reduction in covariate imbalance and lower mean-squared-error in the causal estimate than using either method alone. The improvements occur because the propensity score is a balancing score only asymptotically. In finite samples, some covariate imbalances will remain which another matching method can help adjust.

Matching methods based on the propensity score (estimated by logistic regression), Mahalanobis distance or a combination of the two have appealing theoretical properties if covariates have ellipsoidal distributions—e.g., distributions such as the normal or t. If the covariates are so

[17] The first estimator of treatment effects to be based on a weighted function of the probability of treatment was the Horvitz-Thompson statistic (Horvitz and Thompson 1952).

[18] Optimal matching might sometimes match treated units to non-nearest control units in order to minimize the overall distance (Hansen 2004; Rosenbaum 1991).

distributed, these methods (more generally, affinely invariant matching methods[19]) have the property of "equal percent bias reduction" (EPBR) (Rubin 1976a,b; Rubin and Thomas 1992).[20] This property, which is formally defined in Appendix A, ensures that matching methods will reduce bias in all linear combinations of the covariates. If a matching method is not EPBR, then that method will, in general, increase the bias for some linear function of the covariates even if all univariate means are closer in the matched data than the unmatched (Rubin 1976a).

A significant shortcoming of these common matching methods is that they may (and in practice, frequently do) make balance worse across measured potential confounders. These methods may make balance worse, in practice, even if covariates are distributed ellipsoidally symmetric, because EPBR is a property that holds in expectation. That is, even if the covariates have elliptic distributions, finite samples may not conform to ellipticity, and hence Mahalanobis distance may not be optimal because the matrix used to scale the distances, the sample covariance matrix of X, may not be sufficient for accounting for all of the differences between the distributions of the covariates in X. In finite samples, there may be more differences between the distributions of covariates than just means and variances—e.g., the other moments may differ as well.[21] Moreover, if covariates are neither ellipsoidally symmetric nor are mixtures of DMPES distributions, propensity score matching has good theoretical properties only if the true propensity score model is known with certainty and the sample size is large.

The EPBR property itself is limited and in a given substantive problem it may not be desirable. This can arise if it is known based on theory that one covariate has a large nonlinear relationship with the outcome while another does not—e.g., $Y = X_1^4 + X_2$, where $X > 1$ and where both X_1 and X_2 have the same distribution. In such a case, covariate imbalance in X_1 will be generally more important than X_2 because the response surface (i.e., the model of Y) is more sensitive to changes in X_1 than X_2.

[19] Affine invariance means that the matching output is invariant to matching on X or an affine transformation of X.

[20] The EPBR results of Rubin and Thomas (1992) have been extended by Rubin and Stuart (2006) to the case of discriminant mixtures of proportional ellipsoidally symmetric (DMPES) distributions. This extension is important, but it is restricted to a limited set of mixtures.

[21] On Mahalanobis distance and distributional considerations, see Mitchell and Krzanowski (1985, 1989).

3.2 Genetic Matching

Given these limitations, it may be desirable to use a matching method which algorithmically imposes certain properties when the EPBR property does not hold. One method which does this while keeping the estimand constant is Genetic Matching (GenMatch) (Diamond and Sekhon 2005; Sekhon In Press). Genetic Matching, is a method which automatically finds the set of matches which minimize the discrepancy between the distribution of potential confounders in the treated and control groups—i.e., covariate balance is maximized. GenMatch is a generalization of propensity score and Mahalanobis distance matching. It has been used by a variety of researchers (e.g., Bonney, Canes-Wrone, and Minozzi 2007; Boyd, Epstein, and Martin 2008; Eggers and Hainmueller 2008; Gilligan and Sergenti 2008; Gordon and Huber 2007; Heinrich 2007; Herron and Wand 2007; Korkeamäki and Uuistalo In Press; Lenz and Ladd 2006; Raessler and Rubin 2005; Woo, Reiter, and Karr In Press). The algorithm uses a genetic algorithm (Mebane and Sekhon In Press; Sekhon and Mebane 1998) to optimize balance as much as possible given the data. The method is nonparametric and does not depend on knowing or estimating the propensity score, but the method is improved when a propensity score is incorporated. Diamond and Sekhon (2005) use this algorithm to show that the long running debate between Dehejia and Wahba (2002; 1997; 1999; Dehejia 2005) and Smith and Todd (2005b,a, 2001) is largely a result of researchers using models which do not produce good balance—even if some of the models get close by chance to the experimental benchmark of interest. They show that Genetic Matching is able to quickly find good balance and to reliably recover the experimental benchmark. Sekhon and Grieve (2008) show that for a clinical intervention of interest in the matching literature, Pulmonary Artery Catheterization, applying Genetic Matching to an observational study replicates the substantive results of a corresponding randomized controlled trial unlike the extant literature.

A difficult question which all matching methods must confront in practice is how to measure covariate balance. Users of propensity score matching iterate between tweaking the specification of their propensity score model and then checking the covariate balance. Researchers stop when they are satisfied with the covariate balance they have obtained or when they tire. One process

for cycling between checking for balance on the covariates and reformulating the propensity score model is outlined in Rosenbaum and Rubin (1984). Genetic Matching is an alternative to this process of reformulating the propensity score model, and like other forms of matching it is agnostic about how covariate balance is measured because this is an open research question. Therefore, the GenMatch software (Sekhon In Press) offers a variety of ways to measure covariate balance, many of which rely upon cumulative probability distribution functions. By default, these statistics include paired t-tests, univariate and multivariate Kolmogorov-Smirnov tests. A variety of descriptive statistics based on empirical-QQ plots are also offered. The statistics are not used to conduct formal hypothesis tests, because no measure of balance is a monotonic function of bias in the estimand of interest and because we wish to maximize balance without limit (Imai, King, and Stuart 2008; Sekhon 2006). GenMatch can maximize balance based on a variety of pre-defined measures of balance or any measure the researcher may wish to use such as the Kullback-Leiber divergence measure which is popular in information theory and image processing (Kullback and Leibler 1951). For details see Sekhon (In Press).

4 GOTV Controversy

In a landmark study of various Get-Out-The-Vote (GOTV) interventions, Gerber and Green (2000) reported results from a field experiment in New Haven they conducted in 1998. Revisiting Eldersveld (1956), Gerber and Green examined the relative effectiveness of various GOTV appeals, including short non-partisan telephone calls, direct mail and personal canvassing. They found that "[v]oter turnout was substantially increased by personal canvassing, slightly by direct mail, and not at all by telephone calls" (p653). These results held for both ITT and ETT. The non-compliance problem in this experiment consists of only single-crossover—i.e., there are two types of units, compliers and never-treat. With random assignment of the intention to treat, ETT can be estimated consistently with the 2SLS approach of Equation 4, which Gerber and Green used.

Imai (2005) argued that the attempt to randomly assign treatment in the Green and Gerber study was not successful, and hence, the field experiment should be analyzed using observational methods alone. It was argued that neither ITT nor ETT could be estimated without adjustment. Imai used propensity score matching to estimate ETT. Imai assumed that once a set of observables had been matched upon using his propensity score, the outcomes of compliers assigned to treatment could be compared with the outcomes of units assigned to control to estimate ETT. The inferential problem is that the control group consists of both never-treats and compliers while the units assigned to treatment who received treatment are all compliers.

The observables used by Imai were drawn from the usual voter registration files. There were six covariates for each subject. The indicator variables were: turnout in the prior election, 1996; new voter registrant; major party registrant; and single-voter household. And the two additional covariates were the age of the subject and the ward of residence.

Imai argued that contrary to the original findings, short non-partisan telephone appeals did have a significant positive effect on turnout. Green and Gerber responded in various articles (Arceneaux et al. 2006; Gerber and Green 2005). And Bowers and Hansen entered the debate using alternative methods (Bowers and Hansen 2005; Hansen and Bowers In Press) which reconfirmed the substantive findings of Gerber and Green (2000).

Imai (2005) performed an invaluable service by prompting Gerber and Green to find and correct a number of data-processing errors in the original Gerber and Green (2000) study.[22] Imai also performed an important service by pointing out that at the level of *individuals*, the experiment did not appear to be randomized successfully (even after data-processing errors were corrected)—i.e., covariate imbalances between treatment and control were greater than one would expect by chance. In the original study, the data were analyzed as if individuals were randomized even though households we actually randomized. Prompted by Imai, subsequent randomization checks were performed at the household level once household identifiers were released.

[22] According to Gerber and Green (2005), there were data-processing errors related to: (1) imperfect matches between names on the original master file and the names returned by canvassers; (2) a failure of communication with the phone bank about which treatment groups were to be assigned the GOTV appeal; (3) data manipulation errors that resulted in some subjects in the control group being incorrectly recorded as treatment subjects.

Consistent with the findings of Gerber and Green (2000), all analysts aside from Imai have concluded that short non-partisan telephone calls are not effective. This holds in the original data for the New Haven study (Bowers and Hansen 2005; Gerber and Green 2005), the corrected data (Hansen and Bowers In Press; Gerber and Green 2005) and in subsequent large scale field experiments conducted in Michigan and Iowa (Arceneaux et al. 2006).

This exchange highlights an important lesson: when analyzing any experiment, one should stay as close to the experimental design as possible. This holds even if one conjectures that randomization has not fully balanced the covariates in the given sample. Discarding the experimental design and reverting to purely observational methods fails to result in unbiased estimates of the effectiveness of short non-partisan telephone calls.

Since treatment in the original New Haven experiment was actually randomized at the level of households and not individuals, all randomization checks should be conducted at the household level. Failing to do so results in spuriously finding that randomization failed to balance the observable covariates, when in fact it had. And, ideally, variance estimates should take into account that randomization was done at the level of household as opposed to individuals although in this example this does not appear to make a significant substantive difference because the number of households is large.

With the corrected data, when the randomization checks are performed at the level of household, one finds that randomization was successful (Gerber and Green 2005; Hansen and Bowers In Press). Therefore, no method is needed to correct for any randomization issues. Before the household data was available and before it was known by Imai or Bowers and Hansen that randomization was done at the level of household, it was found that if matching was used to simply strengthen the randomization—i.e., the randomization was not ignored, the original Gerber and Green results were recovered (Bowers and Hansen 2005). The simplest method of strengthening the randomization is to use stratification—i.e., to apply the estimator in Equation 4 within strata defined by observed confounders. Within each stratum, the confounders used to define the strata obviously cannot be an issue (if the covariates are homogeneous within strata).

Even if the original New Haven dataset is examined, and randomization is ignored, Imai's results are not robust to slight changes in methodology such as correcting his biased variance estimates. Unconventionally, Imai reported not the full sample point estimate, but the average estimate from 500 bootstrap estimates. However, using the full sample point estimate results in a $p-$value that is not significant at conventional test levels, even if one uses Imai's bootstrap variance estimate (Gerber and Green 2005). But bootstrapping yields biased variance estimates for matching estimators (Abadie and Imbens 2006b). If one does not use the bootstrap but, for example, the Abadie and Imbens (2006a) approach to estimate the point and variance estimates, one does not obtain a significant estimate at conventional levels.[23] The same holds if one uses Imai's own code but simply does 1-to-1 matching with replacement (Gerber and Green 2005).

Matching in this example fails at least two different placebo tests. Placebo tests are underused as robustness checks in observational studies. Such tests are the observational equivalent of giving a patient a sugar pill in the control group in a clinical trail. We know *a priori* that such a pill should have a zero treatment effect because of our knowledge of the biochemical properties of sugar pills.[24] Therefore, the biochemical effectiveness of the actual treatment of interest can be estimated by comparing it to the results from the placebo group. In an observational placebo test, one attempts to find a strata of data and an outcome for which the treatment effect is known with similar certainty. And then one tests to see if the observational method one is using is able to recover the result which is known *a priori*. In this fashion, one checks both the selection on observables assumption and the estimator at once. In the present case there are two obvious placebo tests.

The first, which is the clearest because it follows directly from the assumptions of the matching estimator, is to estimate the causal effect of being assigned to treatment but never receiving it. Since being assigned to receive a telephone GOTV appeal but never receiving the appeal cannot logically have an effect on turnout, we have a clear placebo. The causal effect must logically be

[23]The point estimate is 5.6 percent, and the Abadie-Imbens standard error is 3.2.

[24]If the placebo does have an effect, we know it cannot be because of any biochemical property of the pill itself. So, the placebo group would still serve as a useful benchmark against which to measure the treatment of interest.

zero. The outcomes of never-treat units who were assigned to treatment are being compared with the outcomes of the never-treat units who were assigned to control. The control group, however, consists of units who would be never-treat if they were assigned to treatment and units who would be compliers. For a valid comparison, one has to find the never-treat in the control group to compare with the never-treat who are assigned to treatment. Imai's observational approach purports to solve this inferential problem since he has to find the compliers in control to compare with the compliers in treatment. Unfortunately, the estimate produced for this placebo test by one-to-five propensity score matching, the type used in Imai (2005), is -5.6 percent with a standard error of 2.3 (Gerber and Green 2005).

A second placebo test is offered by considering if telephone calls have a zero effect on *past* turnout. In this setup, one obviously does not match on previous turnout since that becomes the "outcome" of interest, but one does match on the turnout before the placebo outcome. This placebo test is most appropriate for the Michigan and Iowa experiments described in Arceneaux et al. (2006) because of the availability of turnout history during the past two elections. In these experiments, exact matching estimates ETT to be 1.61 percent with an Abadie-Imbens (Abadie and Imbens 2006a) standard error of 0.258.[25] Exact matching was used to condition on turnout in the election before, age, gender, competitiveness, and household size. As in the previous placebo test, matching claims to find an effect where none is logically possible.

Both of these placebo tests, if conducted, would probably have given any analyst pause. But as is all too common, the selection on observables assumption is accepted readily, by reviewers, readers and most importantly by data analysts themselves. Placebo tests, even when they are possible as in the present case, are rarely conducted.

This behavior is consistent with what has been observed in other disciplines, including economics, epidemiology and clinical medicine. Experimental results are rarely recovered by observational methods, and placebo tests are usually not done, and when they are reported by some researcher to caution against the use of observational methods, such tests are usually ignored. This

[25] Estimated using the Matching package (Sekhon In Press) for the R Project for Statistical Computing.

occurs even in cases where lives are at stake. Note the recent tragic case of Hormone Replacement Therapy (HRT) where tens of thousands of women probably died because their physicians prescribed HRT based on observational studies (Freedman and Petitti 2005a,b).

This GOTV exchange is odd. And its oddity highlights our discipline's belief in models. In order to use a matching algorithm, one need not have discarded all information about the experiment and reverted to purely observational methods. The hybrid approach of Bowers and Hansen (2005) allows one to adjust for any imbalance which remains in the observed covariates while using the information in the randomization. Both this hybrid approach and 2SLS with covariates make the same identification assumption. Both assume that once we condition on X, we can proceed as if the treatment assigned in the experiment is random and as if the compliance model described in the previous section holds. The two methods just differ in how they condition on X: via a parametric model or via stratification or matching. In contrast, as stated before, matching alone makes the same identifying assumption as OLS. Both methods rely upon the selection on observables identification assumption, and they differ in the extent to which they rely upon functional form assumptions.

Given the results of this debate, it is clear that the selection on observables assumption is not valid in this case. And there may be lessons of general interest:

(i) ITT should always be reported, and going beyond ITT should only be done with care.

(ii) All data analysis should leverage the experimental design as much as possible.

(iii) Our belief should be that selection on observables and other identifying assumptions not guaranteed by the design are incorrect unless compelling evidence to the contrary is provided.

(iv) Placebo tests should be conducted whenever possible, and observational studies without them should be marked down.

5 Conclusion

As a discipline, we value novelty. But we do not want to change radically. We like new twists which do not challenge our standard research practices. With both quantitative and qualitative methods, we hope that the next innovation will solve our inference problems. Since we have tried to mass produce science on the cheap, we should not be surprised that a tradition which relies on finding a valid design is not dominant.

These observations are not new. David Freedman has made similar comments over the years about our discipline in particular and the social sciences in general (e.g., Freedman 1995, 1999, 2008a). In one famous example, he contrasted our norms and methods with the case of John Snow and cholera, a prominent example of the success of observational methods for causal inference (Freedman 1991, 1999; Snow 1855; Vinten-Johansen, Brody, Paneth, Rachman, and Rip 2003). As early as the cholera outbreak of 1831–32, the first to reach England, Snow doubted the miasma theory as it applied to cholera. In the outbreak of 1848, he decided to track the progress of the disease, and he was able to find the index case, John Harnold, and document its spread and natural history. In the 1850s, Snow accumulated data on the epidemics of 1853–54 and was able analyze the "grand experiment" which showed that the disease could be linked with specific water suppliers. The Broad Street pump natural experiment occurred in 1854. In 1831, Snow had a hypothesis based on evidence, but no compelling design to make a rigorous causal inference. For a compelling set of natural experiments he had to wait for 1854. More than 20 years had passed since 1831 and six since 1848. A young researcher today who waited that long to find the right design would soon be out of a job. Researchers know this and adapt.

It should be no surprise that the modeling enterprise is the dominant one. Unfortunately, as matching is gaining popularity, its ties to the Neyman-Rubin causal model and considerations of design are weakening. Rubin (2008) notes that "design trumps analysis," but designs for observational data cannot be mass produced. From hunger comes our belief in analysis by models, statistical or otherwise, matching or kernel estimation, maximum likelihood or Bayesian.

For most researchers, the math obscures the assumptions. Without an experiment, a natural

experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive. This conclusion has implications for the kind of causal questions we are able to answer with some rigor. Clear, manipulable treatments and rigorous designs are essential. And the only designs I know of that can be mass produced with relative success rely on random assignment. Rigorous observational studies are important and needed. But I do not know how to mass produce them.

A Equal Percent Bias Reduction (EPBR)

Affinely invariant matching methods, such as Mahalanobis metric matching and propensity score matching (if the propensity score is estimated by logistic regression), are equal percent bias reducing if all of the covariates used have ellipsoidal distributions (Rubin and Thomas 1992)—e.g., distributions such as the normal or t—or if the covariates are mixtures of proportional ellipsoidally symmetric (DMPES) distributions Rubin and Stuart (2006).[26]

To formally define EPBR, let Z be the expected value of X in the matched control group. Then, as outlined in Rubin (1976a), a matching procedure is EPBR if

$$E(X \mid T = 1) - Z \quad = \gamma \{E(X \mid T = 1) - E(X \mid T = 0)\}$$

for a scalar $0 \leq \gamma \leq 1$. In other words, we say that a matching method is EPBR for X when the percent reduction in the biases of each of the matching variables is the same. One obtains the same percent reduction in bias for any linear function of X if and only if the matching method is EPBR for X. Moreover, if a matching method is not EPBR for X, the bias for some linear function of X is increased even if all univariate covariate means are closer in the matched data than the unmatched (Rubin 1976a).

References

Abadie, Alberto and Guido Imbens. 2006a. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74: 235–267.

Abadie, Alberto and Guido Imbens. 2006b. "On the Failure of the Bootstrap for Matching Estimators." Working Paper.

[26]Note that DMPES defines a limited set of mixtures. In particular, countably infinite mixtures of ellipsoidal distributions where: (1) all inner products are proportional and (2) where the centers of each constituent ellipsoidal distribution are such that all best linear discriminants between any two components are also proportional.

Angrist, Joshua D. and Guido W. Imbens. 1994. "Identication and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–475.

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.

Anscombe, F. J. 1948. "The Validity of Comparative Experiments." *Journal of the Royal Statistical Society, Series A* 61: 181–211.

Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political Analysis* 14 (1): 37–62.

Armitage, Peter. 2003. "Fisher, Bradford Hill, and Randomization." *International Journal of Epidemiology* 32 (6): 925–928.

Barnard, John, Constantine E. Frangakis, Jennifer L. Hill, and Donald B. Rubin. 2003. "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City." *Journal of the American Statistical Association* 98 (462): 299–323.

Bloom, Howard S. 1984. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review* 8 (2): 225–246.

Bonney, Jessica, Brandice Canes-Wrone, and William Minozzi. 2007. "Issue Accountability and the Mass Public: The Electoral Consequences of Legislative Voting on Crime Policy." Working Paper.

Bowers, Jake and Ben Hansen. 2005. "Attributing Effects to a Get-Out-The-Vote Campaign Using Full Matching and Randomization Inference." `http://www-personal.umich.edu/~jwbowers/PAPERS/bowershansen03Apr05.pdf`.

Boyd, Christina L. Lee Epstein, and Andrew D. Martin. 2008. "Untangling the Causal Effects of Sex on Judging." 2nd Annual Conference on Empirical Legal Studies Paper. Available at SSRN: http://ssrn.com/abstract=1001748.

Brady, Henry. 2008. "Causation and Explanation in Social Science." In Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, editors, *The Oxford Handbook of Political Methodology* New York: Oxford University Press. pages 217–270.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and R.A. Olshen. 1984. *Classification and Regression Trees*. New York: Chapman & Hall.

Chapin, Stuart F. 1938. "Design for Social Experiments." *American Sociological Review* 3 (6): 786–800.

Chipman, Hugh A. Edward I. George, and Robert E. McCulloch. 2006. "BART: Bayesian Additive Regression Trees." Working Paper.

Christakis, Nicholas A. and Theodore I. Iwashyna. 2003. "The Health Impact of Health Care on Families: A matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses." *Social Science & Medicine* 57 (3): 465–475.

Cochran, William G. 1953. "Matching in Analytical Studies." *American Journal of Public Health* 43: 684–691.

Cochran, William G. 1965. "The Planning of Observational Studies of Human Populations (with discussion)." *Journal of the Royal Statistical Society, Series A* 128: 234–255.

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

Cox, David R. 1958. *Planning of Experiments*. New York: Wiley.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2006. "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand." Working Paper.

Dehejia, Rajeev. 2005. "Practical Propensity Score Matching: A Reply to Smith and Todd." *Journal of Econometrics* 125 (1–2): 355–364.

Dehejia, Rajeev and Sadek Wahba. 1997. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." Rejeev Dehejia, *Econometric Methods for Program Evaluation*. Ph.D. Dissertation, Harvard University, Chapter 1.

Dehejia, Rajeev and Sadek Wahba. 1999. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–1062.

Dehejia, Rajeev H. and Sadek Wahba. 2002. "Propensity Score Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84 (1): 151–161.

Diamond, Alexis and Jasjeet S. Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." Working Paper.

Diprete, Thomas A. and Henriette Engelhardt. 2004. "Estimating Causal Effects With Matching Methods in the Presence and Absence of Bias Cancellation." *Sociological Methods & Research* 32 (4): 501–528.

Dorn, H. F. 1953. "Philosophy of Inference from Retrospective Studies." *American Journal of Public Health* 43: 692–699.

Dunning, Thad. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Science Quarterly* 61 (2): 282–293.

Eggers, Andy and Jens Hainmueller. 2008. "The Value of Political Power: Estimating Returns to Office in Post-War British Politics." Working Paper.

Eldersveld, Samuel J. 1956. "Experimental Propaganda Technique and Voting Behavior." *American Political Science Review* 50 (1): 154–165.

Fechner, Gustav Theodor. 1966 [1860]. *Elements of psychophysics, Vol 1.*. New York: Rinehart & Winston. Translated by Helmut E. Adler and edited by D.H. Howes and E.G. Boring.

Freedman, David A. 1991. "Statistical Models and Shoe Leather." *Sociological Methodology* 21: 291–313.

Freedman, David A. 1995. "Some Issues in the Foundation of Statistics." *Foundations of Science* 1: 19–39.

Freedman, David A. 1999. "From Association to Causation: Some Remarks on the History of Statistics." *Statistical Science* 14: 243–258.

Freedman, David A. 2004. "On Specifying Graphical Models for Causation, and the Identification Problem." *Evaluation Review* 26 (4): 267–93.

Freedman, David A. 2006. "Statistical Models for Causation: What Inferential leverage Do They Provide?" *Evaluation Review* 30: 691–713.

Freedman, David A. 2008a. "Oasis or Mirage?" *CHANCE Magazine* 21 (1): 59–61.

Freedman, David A. 2008b. "On Regression Adjustments in Experiments with Several Treatments." *Annals of Applied Statistics* 2 (1): 176–196.

Freedman, David A. 2008c. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics* 40 (2): 180–193.

Freedman, David A. 2008d. "Randomization Does not Justify Logistic Regression." *Statistical Science* 23 (2): 237–249.

Freedman, David A. and Diana B. Petitti. 2005a. "Hormone Replacement Therapy Does Not Save Lives: Comments on the Women's Health Initiative." *Biometrics* 61 (4): 918–920.

Freedman, David A. and Diana B. Petitti. 2005b. "Invited Commentary: How Far Can Epidemiologists Get with Statistical Adjustment?" *American Journal of Epidemiology* 162 (5): 415–418.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky. 2005. "Water for Life: The Impact of the Privatization of Water Services on Child Mortality." *Journal of Political Economy* 113 (1): 83–120.

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94 (3): 653–663.

Gerber, Alan S. and Donald P. Green. 2005. "Correction to Gerber and Green (2000) Replication of Disputed Findings, and Reply to Imai (2005)." *American Political Science Review* 99 (2): 301–313.

Gilligan, Michael J. and Ernest J. Sergenti. 2008. "Evaluating UN Peacekeeping with Matching to Improve Causal Inference." *Quarterly Journal of Political Science* 3 (2): 89–122.

Gordon, Sandy and Greg Huber. 2007. "The Effect of Electoral Competitiveness on Incumbent Behavior." *Quarterly Journal of Political Science* 2 (2): 107–138.

Gosnell, Harold F. 1927. *Getting Out the Vote: an experiment in the stimulation of voting*. Chicago: University of Chicago Press.

Gosnell, Harold F. 1948. "Mobilizing the Electorate." *Annals of the American Academy of Political and Social Science* 259: 98–103.

Green, Donald and Alan Gerber. 2002. "Reclaiming the Experimental Tradition in Political Science." In Helen Milner and Ira Katznelson, editors, *State of the Discipline, Vol. III* New York: W.W. Norton & Company, Inc. pages 805–832.

Greenwood, Ernest. 1945. *Experimental Sociology: A Study in Method*. New York: King's Crown Press.

Haavelmo, Trygve. 1943. "The Statistical Implications of a System of Simultaneous Equations." *Econometrica* 1 (11): 1–12.

Hall, Nancy. 2007. "R. A. Fisher and His Advocacy of Randomization." *Journal of the History of Biology* 40 (2): 295–325.

Hansen, Ben B. 2004. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association* 99: 609–618.

Hansen, Ben B. and Jake Bowers. In Press. "Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign." *Journal of the American Statistical Association*.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66 (5): 1017–1098.

Heinrich, Caryolyn J. 2007. "Demand and Supply-Side Determinants of Conditional Cash Transfer Program Effectiveness." *World Development* 35 (1): 121–143.

Herron, Michael C. and Jonathan Wand. 2007. "Assessing Partisan Bias in Voting Technology: The Case of the 2004 New Hampshire Recount." *Electoral Studies* 26 (2): 247–261.

Hill, Bradford. 1961. *Principles of Medical Statistics*. London: The Lancet 7 edition.

Hodges, J. L and Erich L. Lehmann. 1964. *Basic Concepts of Probability and Statistics*. San Fransisco: Holden-Day, Inc.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–960.

Horvitz, D. G. and D. J. Thompson. 1952. "A Generalization of Sampling without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47: 663–685.

Imai, Kosuke. 2005. "Do Get-Out-The-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review* 99 (2): 283–300.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. "Misunderstandings among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 171 (2): 481–502.

Imai, Kosuke and David A. van Dyk. 2004. "Causal Inference With General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99 (467): 854–866.

Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87 (3): 706–710.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86 (1): 4–29.

Imbens, Guido W. and Jeffrey M. Wooldridge. 2008. "Recent Developments in the Econometrics of Program Evaluation." NBER Working Paper No. 14251.

Kempthorne, O. 1952. *The Design and Analysis of Experiments*. New York: Wiley.

Kempthorne, O. 1955. "The Randomization Theory of Experimental Inference." *Journal of the American Statistical Association* 50: 495–497.

Korkeamäki, Ossi and Roope Uuistalo. In Press. "Employment and Wage Effects of a Payroll-Tax Cut—evidence from a regional experiment." *International Tax and Public Finance*.

Kullback, Solomon and Richard A. Leibler. 1951. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22: 79–86.

Lee, David S. 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics* 142 (2): 675–697.

Lehmann, Erich L. 1993. "The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?" *Journal of the American Statistical Association* 88 (424): 1242–1249.

Lenz, Gabriel S. and Jonathan McDonald Ladd. 2006. "Exploiting a Rare Shift in Communication Flows: Media Effects in the 1997 British Election." `http://sekhon.berkeley.edu/causalinf/papers/LaddLenzBritish.pdf`.

Lewis, David K. 1973. *Counterfactuals*. Cambridge, MA: Harvard University Press.

Mauldon, Jane, Jan Malvin, Jon Stiles, Nancy Nicosia, and Eva Seto. 2000. "Impact of California's Cal-Learn Demonstration Project: Final Report." UC DATA Archive and Technical Assistance.

McCarthy, M. D. 1939. "On the Application of the z-Test to Randomized Blocks." *Ann. Math. Statist.* 10: 495–497.

Mebane, Walter R. Jr. and Jasjeet S. Sekhon. In Press. "Genetic Optimization Using Derivatives: The `rgenoud` package for `R`." *Journal of Statistical Software*.

Middleton, Joel A. 2008. "Bias of the Regression Estimator for Experiments using Clustered Random Assignment." *Statistics & Probability Letters* 78 (16): 2654–2659.

Mitchell, Ann F. S. and Wojtek J. Krzanowski. 1985. "The Mahalanobis Distance and Elliptic Distributions." *Biometrika* 72 (2): 464–467.

Mitchell, Ann F. S. and Wojtek J. Krzanowski. 1989. "Amendments and Corrections: The Mahalanobis Distance and Elliptic Distributions." *Biometrika* 76 (2): 407.

Morgan, Stephen L. and David J. Harding. 2006. "Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice." *Sociological Methods & Research* 35 (1): 3–60.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Mueller, Franz H. 1945. "Review of: "Experimental Sociology: A Study in Method" by Ernest Greenwood." *The American Catholic Sociological Review* 6 (3): 185–186.

Neyman, Jerzy. 1923 [1990]. "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.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Pitman, E. J. G. 1937. "Significance Tests which can be Applied to Samples from any Populations. III. The Analysis of Variance Test." *Biometrika* 29: 322–335.

Preece, D. A. 1990. "R. A. Fisher and Experimental Design: A Review." *Biometrics* 46 (4): 925–935.

Raessler, S. and D. B. Rubin. 2005. "Complications when using nonrandomized job training data to draw causal inferences." *Proceedings of the International Statistical Institute*.

Reid, Constance. 1982. *Neyman from Life*. New York: Springer.

Rosenbaum, Paul R. 1991. "A Characterization of Optimal Designs for Observational Studies." *Journal of the Royal Statistical Society, Series B* 53 (3): 597–610.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer-Verlag 2nd edition.

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

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rosenbaum, Paul R. and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (387): 516–524.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician* 39 (1): 33–38.

Rosenzweig, Mark R. and Kenneth I. Wolpin. 2000. "Natural "Natural Experiments" in Economics." *Journal of Economic Literature* 38 (December): 827–874.

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

Rubin, Donald B. 1976a. "Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples." *Biometrics* 32 (1): 109–120.

Rubin, Donald B. 1976b. "Multivariate Matching Methods That are Equal Percent Bias Reducing, II: Maximums on Bias Reduction for Fixed Sample Sizes." *Biometrics* 32 (1): 121–132.

Rubin, Donald B. 1977. "Assignment to a Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2: 1–26.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6 (1): 34–58.

Rubin, Donald B. 1979. "Using Multivariate Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74: 318–328.

Rubin, Donald B. 1980. "Bias Reduction Using Mahalanobis-Metric Matching." *Biometrics* 36 (2): 293–298.

Rubin, Donald B. 1990. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science* 5 (4): 472–480.

Rubin, Donald B. 1997. "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Annals of Internal Medicine* 127 (8S): 757–763.

Rubin, Donald B. 2001. "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation." *Health Services & Outcomes Research Methodology* 2 (1): 169–188.

Rubin, Donald B. 2006. *Matched Sampling for Causal Effects*. New York: Cambridge University Press.

Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2 (3): 808–840.

Rubin, Donald B. and Elizabeth A. Stuart. 2006. "Affinely Invariant Matching Methods with Discriminant Mixtures of Proportional Ellipsoidally Symmetric Distributions." *Annals of Statistics* 34 (4): 1814–1826.

Rubin, Donald B. and Neal Thomas. 1992. "Affinely Invariant Matching Methods with Ellipsoidal Distributions." *Annals of Statistics* 20 (2): 1079–1093.

Scheffé, H. 1956. "Alternative Models for the Analysis of Variance." *Annals of Mathematical Statistics* 27: 251–271.

Sekhon, Jasjeet S. 2004. "The Varying Role of Voter Information Across Democratic Societies." Working Paper.
URL `http://sekhon.berkeley.edu/papers/SekhonInformation.pdf`

Sekhon, Jasjeet S. 2006. "Alternative Balance Metrics for Bias Reduction in Matching Methods for Causal Inference." Working Paper.
URL `http://sekhon.berkeley.edu/papers/SekhonBalanceMetrics.pdf`

Sekhon, Jasjeet S. In Press. "Matching: Multivariate and Propensity Score Matching with Automated Balance Search." *Journal of Statistical Software*. Computer program available at `http://sekhon.berkeley.edu/matching/`.

Sekhon, Jasjeet S. and Richard Grieve. 2008. "A New Non-Parametric Matching Method for Bias Adjustment with Applications to Economic Evaluations." Working Paper.

Sekhon, Jasjeet Singh and Walter R. Mebane, Jr. 1998. "Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models." *Political Analysis* 7: 189–203.

Simon, Herbert. 1953. "Causal Ordering and Identifiability." In William C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method* New York: Wiley. pages 49–74.

Smith, Herbert L. 1997. "Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology* 27: 305–353.

Smith, Jeffrey and Petra Todd. 2005a. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1–2): 305–353.

Smith, Jeffrey and Petra Todd. 2005b. "Rejoinder." *Journal of Econometrics* 125 (1–2): 365–375.

Smith, Jeffrey A. and Petra E. Todd. 2001. "Reconciling Conflicting Evidence on the Performance of Propensity Score Matching Methods." *AEA Papers and Proceedings* 91 (2): 112–118.

Snow, John. 1855. *On the Mode of Communication of Cholera*. London: John Churchill 2nd edition.

Sommer, Alfred and Scott L. Zeger. 1991. "On Estimating Efficacy from Clinical Trials." *Statistics in Medicine* 10 (1): 45–52.

Speed, Terence P. 1990. "Introductory Remarks on Neyman (1923)." *Statistical Science* 5 (4): 463–464.

Stigler, Stephen M. 1990. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Belknap Press.

Stock, James H. and Francesco Trebbi. 2003. "Who Invented Instrumental Variable Regression?" *Journal of Economic Perspectives* 17 (3): 177–194.

Thistlethwaite, Donald L. and Donald T. Campbell. 1960. "Regression-Discontinuity Analysis: An alternative to the ex post facto experiment." *Journal of Educational Psychology* 51 (6): 309–317.

Vinten-Johansen, Peter, Howard Brody, Nigel Paneth, Stephen Rachman, and Michael Russell Rip. 2003. *Cholera, Chloroform, and the Science of Medicine: A Life of John Snow*. New York: Oxford University Press.

Wand, Jonathan N. 2008. "Harold F. Gosnell and Social Science Experiments." Working Paper.

Welch, B. L. 1937. "On the z-Test in Randomized Blocks and Latin Squares." *Biometrika* 29: 21–52.

Winship, Christopher and Stephen Morgan. 1999. "The estimation of causal effects from observational data." *Annual Review of Sociology* 25: 659–707.

Woo, Mi-Ja, Jerome P. Reiter, and Alan F. Karr. In Press. "Estimation of propensity scores using generalized additive models." *Statistics in Medicine*.

Wright, Philip G. 1928. *The Tariff on Animal and Vegetable Oils*. New York: Macmillan.

Yule, Undy G. 1899. "An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades (Part I.)." *Journal of the Royal Statistical Society* 62 (2): 249–295.