

Matching in Randomized Experiments*

Luke Keele

Corrine McConnaughy

Ismail White

July 6, 2008

Abstract

Randomization in experiments allows researchers to assume that the treatment and control groups are balanced with respect to all characteristics except the treatment. Randomization, however, only makes balance probable, and accidental covariate imbalance can occur for any specific randomization. As such, statistical adjustments for accidental imbalance are common with experimental data. The most common method of adjustment for accidental imbalance is to use least squares to estimate the analysis of covariance (ANCOVA) model. ANCOVA, however, is a poor choice for the adjustment of experimental data. It has a strong functional form assumption, and the least squares estimator is notably biased in sample sizes of less than 500 when applied to the analysis of treatment effects. We evaluate alternative methods of adjusting experimental data. We compare ANCOVA to two different techniques. The first technique is a modified version of ANCOVA that relaxes the strong functional form assumption of this model. The second technique is matching, and we test the differences between two matching methods. For the first, we match subjects and then randomize treatment across pairs. For the second, we randomize the treatment and match prior to the estimation of treatment effects. We use all three techniques with data from a series of experiments on racial priming. We find that matching substantially increases the efficiency of experimental designs.

1 Introduction

Researchers looking to assess a causal inference have come to regard the randomized experiment as “the gold standard” (see Bowers and Hansen (2007)) of their enterprise. The gold standard nomenclature results from the promise of randomization to produce equivalence of a treatment and control group, with the exception of receipt of the treatment itself, enabling confident statements about the internal validity of inferences about treatment effects. That is, randomization is credited with the unique ability to induce sameness across treatment and control groups in both their pre-treatment levels on the dependent variable and their propensity to respond to the treatment. The latter is generally thought of as a function of the

*Authors are in alphabetical order.

characteristics that induce a treatment response - what are generally termed mediating and moderating characteristics (variables) - though it may also be a function of the dependent variable, itself (i.e., responsiveness may depend on initial levels of the dependent variable).

How well the meaning behind the “gold standard” label applies to any one experiment, however, is dependent upon a number of factors. First and foremost, because randomization is a tool that delivers equivalence of treatment and control groups in expectation only, by chance any one experiment - run once - stands threatened by the possibility that the groups are not equivalent. This factor is a purely random error, of course. Other factors that hinder a researchers ability to validly detect a causal effect would include concerns such as measurement error, lack of compliance, and reactivity. Good design seeks to minimize all of these errors as much as possible (under constraints, of course). Statistical techniques might help clean up some of what the design could not. We concentrate in this paper on chance assignment as a threat to valid causal inferences in particular experiments.

That chance assignment might threaten their ability to see clearly the treatment effect (or lack thereof) within their subjects is an established concern among applied experimental researchers. Upon noticing that some characteristics of their subjects are unevenly distributed across their treatment and control groups, experimentalists often attempt to analytically “control for” this imbalance through statistical adjustment. In this paper, we address the experimentalists concerns by laying out clearly the consequences (or lack thereof) of imbalances of different sorts in randomized experiments, importantly delineating a loss of balance from a loss of equivalence.

We then review the design and statistical strategies that might be employed to address the consequences a loss of equivalence. We are careful to point out the assumptions that underlie use-in-practice of each alternative, differentiating between strategies that seem flawed in most or all experimental contexts from those that may be more comfortably justified under certain conditions. In particular we contrast model based strategies with design based strategies. We also propose matching as an alternative to model based strategies when design based adjustment is impractical. We illustrate and assess the differences between these strategies using both an original experiment on priming and existing data from an experiment on levels of support for Iraq. Generally, the arguments that follow are orthogonal to whether one

chooses to analyze experiments using either the randomization test framework of Fisher or the average treatment effects framework of Neyman. Either method is compatible with the strategies of adjustment that we compare.

2 Equivalence, Treatment Effects, and Covariates

The randomized experiment is attributed to the work of Fisher at the Rothamstead agricultural station and is expounded in his seminal work “Design of Experiments” (1935). The formal framework for experiments, however, originated with Neyman (1923). Under the Neyman model, each unit under study has two potential outcomes, one if the unit receives a stimulus or treatment and another if the unit remains untreated. Neyman defines causal effects as the difference between these two potential outcomes. The dilemma is that only one of these two outcomes can ever be observed. More formally, let Y_{i1} be the potential outcome for unit i if the unit receives the treatment, and let Y_{i0} be the potential outcome if the unit is in the control group. Which outcome is observed is determined by a random treatment assignment mechanism, T_i , which takes the value of one if the subject receives the treatment and zero if it does not. The actual outcomes are often written as:

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

The individual level treatment effect for unit i is defined by

$$\Delta_i = Y_{i1} - Y_{i0}.$$

The estimation of treatment effect is a missing data problem since we never observe both Y_{i1} and Y_{i0} . Because Δ_i is by definition unobservable, an estimation strategy must be employed. It is typical first to aggregate over the individual level treatment effects and reconceive the quantity of interest as the sample average treatment effect (SATE) among n subjects. The formal definition of SATE is:

$$SATE = \frac{1}{n} \sum_{i=1}^n \Delta_i$$

A simple, consistent estimator for SATE is the observed difference between the control group and the treatment group in their average post-treatment values on the outcome variable Y_i . If m of n subjects in the experiment received the treatment, then the estimator is defined as:

$$\begin{aligned}\hat{\Delta} &= \frac{1}{m} \sum_{i|T_i=1} Y_{i1} - \frac{1}{n-m} \sum_{i|T_i=0} Y_{i0} \\ &= \bar{y}_1 - \bar{y}_0\end{aligned}\tag{1}$$

As of yet, we have seen no role for characteristics of subjects other than their treatment status and their observed outcome. These other characteristics, often thrown under the umbrella label of “covariates, however, are exactly what concern the experimentalist when she notices that they are unevenly distributed across the treatment and control subjects in her study. To clarify whether, how, and to what extent such concerns are justified, we distinguish between classes of such characteristics based on the type of problems they pose for the experimenters ability to make the correct causal inference from the observed data. Note that any of these characteristics may be observed or unobserved quantities in the actual experimental setting.¹

Of most central concern to the experimenter are typically characteristics that actually (in truth) affect subjects propensity to respond to the treatment. These are characteristics that typically fall under the label “moderating variables.” In other words, these are characteristics that put the researcher into the realm of heterogeneous treatment effects, where Δ_i is not the same for all i , but is variable across i based on levels of the moderating variable. Call such characteristics u . Since moderating variables only affect Y_i in the presence of the treatment, the variation in Δ_i is a function of Y_{i1} only. This implies that

$$Y_{i1} = Y_{i0} + f(u).$$

It is straightforward to see that imbalance in u can induce error in Δ as an estimate of

¹Note that while Imai, King, and Stuart (2008) address the question of covariate imbalance in a single experiment, they do not distinguish between types of covariates in the way we do. We argue that there is real added utility in making such distinctions, both conceptually and practically.

SATE because the equivalence assumption is violated by an unequal propensity to respond to the treatment across the treatment and control groups. The average outcome under treatment that is observed for the treatment group is different from the average outcome under treatment that would have been observed for the control group if they had been treated by the difference in $f(u)$. Thus the estimation error is given by

$$\text{SATE} - \Delta = \frac{\sum_{i=n-m}^n f(u_i) - \sum_{i=1}^m f(u_i)}{n} \quad (2)$$

This type of estimation error is illustrated in Table 1. The table gives the details of an experiment on ten subjects five treated and five not treated. Across the groups there is an imbalance in a possible moderating variable, u . The fifth and sixth columns consider the outcome of the experiment when the treatment effect is not moderated by u , but is a constant one unit difference between Y_1 and Y_0 for all subjects. SATE and Δ are identical despite the “covariate” imbalance. In contrast, the last two columns demonstrate that when the treatment effect is moderated by u , such that the treatment effect is a one unit difference between Y_1 and Y_0 for those subjects who have the moderating characteristic and no difference for those subjects who do not, SATE and $\hat{\Delta}$ diverge by the amount given by Equation 2.

Table 1: Estimation Error Due to Moderator Imbalance

Subject	T_i	u	Y_0	$Y_1 \neq f(u)$	$\Delta \neq f(u)$	$Y_1 = f(u)$	$\Delta = f(u)$	
1	1	1	1	2	1	2	1	
2	1	0	1	2	1	1	0	
3	1	0	1	2	1	1	0	
4	1	0	2	3	1	2	0	
5	1	1	2	3	1	3	1	
6	0	1	1	2	1	2	1	
7	0	1	1	2	1	2	1	
8	0	1	1	2	1	2	1	
9	0	1	2	3	1	3	1	
10	0	1	2	3	1	3	1	
SATE:					1	SATE:		.7
$\hat{\Delta}$:					1	$\hat{\Delta}$:		.4

Of course the equivalence assumption might also be violated by an (unobservable) imbalance in Y_{0i} . The error in Δ as an estimate of SATE is simply the difference between the

treatment and control groups averages on Y_0 :

$$\text{SATE} - \hat{\Delta} = \sum_{i=n-m}^n Y_{0i} - \sum_{i=1}^m Y_{0i} \quad (3)$$

Table 2 illustrates the consequence of an imbalance in Y_0 across treatment and control groups. As in Table 1, we use a ten-subject experiment set-up, where half the subjects are assigned to each condition, and we alternately consider a treatment effect that is and is not moderated by u . This time the fifth and sixth columns highlight the difference between SATE and $\hat{\Delta}$ introduced by the unequal distribution of Y_0 when Δ is not moderated by u , which can be derived from Equation 3. In the final two columns, Table 2 also illustrates what can happen when Δ is moderated by u and both Y_0 and u are imbalanced. In this case, the two sources of estimation error offset each other, and hence the “greater” imbalance actually results in lesser estimation error.

Table 2: Estimation Error Due to Imbalance in Potential Outcome Absent Treatment

Subject	T_i	u	Y_0	$Y_1 \neq f(u)$	$\Delta \neq f(u)$	$Y_1 = f(u)$	$\Delta = f(u)$	
1	1	1	1	2	1	2	1	
2	1	0	1	3	1	2	0	
3	1	0	1	3	1	2	0	
4	1	0	2	3	1	2	0	
5	1	1	2	3	1	3	1	
6	0	1	1	2	1	2	1	
7	0	1	1	2	1	2	1	
8	0	1	1	2	1	2	1	
9	0	1	1	2	1	2	1	
10	0	1	2	3	1	3	1	
SATE:					1	SATE:		.7
$\hat{\Delta}$:					1.6	$\hat{\Delta}$:		1

Although imbalances in u and Y_0 are uniquely responsible for violations of the equivalence assumption in an experiment, experimenters may nonetheless notice and concern themselves with imbalances in other subject characteristics. In order to consider the basis for such concern, we translate our treatment of estimation error into a regression framework. We do so simply for ease of exposition. We begin by outlining the usual statistical model for estimation of Δ , the treatment effect. The term T_i is an indicator for whether a unit has

randomly received the treatment or not. We estimate the treatment effect with least squares using the following regression model

$$Y_i = \beta_0 + \beta_1 T_i \quad (4)$$

where Y_i is the observed outcome for the units with $i = 1, \dots, N$. Our estimate for Δ in this model is $\hat{\beta}_1$. The least squares estimate for β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y}) T_i}{\sum_{i=1}^n (T_i - \bar{T})^2} \quad (5)$$

Since T_i is an indicator variable, its mean is:

$$\bar{T} = \frac{1}{N} \sum_{i=1}^N T_i = p \quad (6)$$

or the observed proportion of the sample that receives the treatment. For one possible application of the treatment, however, it is possible that one covariate is not accounted for by randomization. Assume there is an additional set of covariates that may affect the level of Y_i . We would now write the equation as

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i. \quad (7)$$

If X_i is omitted from the model, we estimate $\tilde{\beta}_1$ as:

$$\tilde{\beta}_1 = (T_i' T_i)^{-1} T_i' Y_i$$

If we take expectations of this estimate, we find:

$$E[\tilde{\beta}_1] = \beta_1 + (T_i' T_i)^{-1} T_i' X_i \beta_2$$

If accidental covariate imbalance occurs then $Cov(T_i, X_i) \neq 0$ and assuming that $\beta_2 \neq 0$ which results in a biased estimate of the treatment effect β_1 . This is equivalent to when there is imbalance in Y_0 as in Table 2. For the derivation below, we assume a constant treatment effect. A moderator could be introduced by adding an interaction between X_i and

T_i in Equation 11.

We can write the bias in the estimate of the treatment effect as

$$\tilde{\beta}_1 - \beta_1 = \frac{\beta_2 \sum (T_i - \bar{T})(X_i - \bar{X})}{\sum (T_i - \bar{T})^2} + \frac{\sum (T_i - \bar{T})}{\sum (T_i - \bar{T})^2} \quad (8)$$

By definition in Equation 8 $\bar{e} = 0$. If we take the probability limit for this above equation, it follows that $\bar{d} \rightarrow 0$ as $n \rightarrow \infty$ and we can rewrite Equation 8 as:

$$\text{plim}(\tilde{\beta}_1 - \beta_1) = \frac{\beta_2 \sum d_i X_i}{n} + \frac{\sum d_i}{n} \quad (9)$$

Asymptotically the bias due to covariate imbalance is:

$$\text{plim}(\tilde{\beta}_1 - \beta_1) = \frac{\beta_2 \sum d_i X_i}{n} \approx \bar{X}_T - \bar{X}_C \quad (10)$$

The bias due to an unbalanced covariate is simply the difference in means across the treatment and control group on the unbalanced covariate. If the imbalance in X_i actually captures the imbalance in Y_0 then this is equivalent to the bias in Equation 3.

Now consider another situation where there is no imbalance in X_i , and therefore $Cov(T_i, X_i) = 0$, but $\beta_2 \neq 0$ still holds. If this is true it is trivial to demonstrate that the treatment effect estimate will be unbiased. However, including X_i in the estimating equation will cause the following to be true: $\sigma_{\tilde{\beta}}^2 < \sigma_{\beta}^2$. Thus including X_i on the right hand side of the model will result in a more precisely estimated treatment effect. This reduces the signal to noise ratio in Y_i enabling the analyst to more clearly detect the treatment effect. However, if $\beta_2 = 0$ this increase the variance in the model creating a less precision in the estimate of the treatment effect. In other words, including an irrelevant covariate makes it more difficult for the analyst to detect the treatment effect.

In the practical realm of experimentation, one additional complication concerning covariates may occur. The experimenter might use covariates in two additional ways. To demonstrate this problem, we introduce some additional notation. Assume that the experimenter observes that some subject characteristics V_i that are thought to be covariates related to Y_i in the following fashion

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \beta_3 V_i. \tag{11}$$

In truth, however, $\beta_3 = 0$. That is these characteristics, V_i , are orthogonal to the outcomes conditional on X_i . In such contexts, experimenters might face two difficulties which result in the same set of problems. First, X_i is unobserved but the analyst fears it is related to Y_i and uses V_i as a proxy for X_i . Or the experimenter notices an imbalance on V_i and assumes that $Cov(T_i, V_i) \neq 0$ and $\beta \neq 0$ thus including it in the equation since X_i is again unobserved. In both instances, if V_i is related to X_i then biased estimates for the treatment effect will occur. Here analytics are unable to give us any insight into the direction of the bias only that it will be present. Both problems are akin to the usual measurement error problems in regression contexts.

Having highlighted the various difficulties that can be presented by imbalance, causal heterogeneity, inefficiency, and equivalence violations in a single experiment, we have opened the question of whether and how one adjust design or analytic strategies to cope with these problems. While there are viable strategies for dealing with these complications, there are often additional assumptions that must be made and numerous practical limitations.

3 Adjusting Experimental Data

We now turn to coping with the myriad of complications that can arise in experimental data. In what follows, we distinguish between design based strategies and analytic strategies. Analytic strategies come in two flavors: one that is model based and another that is not. We should note in advance that often these strategies are not mutually exclusive and may in fact be used in tandem to address different complications. While we acknowledge the limitations of all of these approaches, we pay particular attention to critiques of the model based approach to experimental data.

3.1 Blocking

Blocking is one design based alternative to a model based analysis. In a traditional block design, the analyst stratifies the randomization along levels of some possibly confounding or

unbalanced factor. If we expect an imbalance on race, for example, randomization would be implemented within stratified groups or blocks of racial categories. This is but one type of block design of which there many. Some authors have recently called for greater usage of block designs in experimental research (Imai, King, and Stuart 2008). The drawback to block designs is that they require knowing in advance the covariate that might cause imbalance. If the research is able to identify a strong confounder in advance of data collection, block designs are an excellent strategy. In some situations, this is feasible, but often it is not. Moreover, even if one uses a blocking design, blocking does not prevent imbalance within levels of the blocking factor. Though such imbalance can be corrected with some of the techniques that we discuss later.

Advanced search algorithms can alleviate some of the problems in traditional block designs. Here, the analyst can use a greedy search algorithm to form blocks based on a multivariate measure of distance based on a variety of covariates that might be imbalanced (Moore 2008.). These algorithms are especially useful for what are often called matched pair designs. In a matched pair design, experimental units placed into pairs and the randomization is conducted within the matched pairs. Search algorithms make this matching process much less ad hoc and allow for creating matched pairs based on a number of covariates. The matched pairs design is simply a form of a block design where each block contains only two units. This new form of blocking based on search algorithms has much in common with the matching techniques that have become popular with observational data. Search based blocking shares one flaw with traditional block designs a priori knowledge of the experimental units. This often unproblematic for field experiments where the units based on geography, but for human subjects prescreening can complicate the practical aspects of the experiment. Such blocking requires at least one session for prescreening and then a second session for the actual experiment. With two session questions of compliance can arise. Moreover, if one wishes to use an adult convenience sample one must recruit the subjects, prescreen, and then conduct the experiment. This may prove difficult without offering the subjects a financial incentive.

3.2 Regression Adjustment

The analysis of data in political science is typically model based. Most often some form of regression (linear, logistic, count, etc.) is used to draw conclusions about possible relationships between covariates. Regression models of any form require a series of assumptions about the data generating process and often rely on an asymptotic approximation. Depending on the context, these assumptions and approximations may be quite reasonable or at times may stretch credulity. Model based data analysis is valid only if these assumptions are met and the asymptotic approximation holds. Such model based approaches to data analysis have proven useful in many areas of political science, and the model based approach is often unavoidable with many forms of observational data.

One can use a model based approach to analyze experimental data, but experimental data often do not require the assumptions and approximations needed for regression based models. In fact, we would argue that model based approaches violate the the spirit of experimental research which allows analysts to avoid the assumptions that are part and parcel of model based analysis. With experimental data, the analyst can instead adopt a design based approach, where the experiment is designed to account for possible threats to inference. Within design-based approaches one can use either the exact method of Fisher (1935) or the average treatment effects approach of Neyman (1923). While analysts often differ over which approach is more suitable, both methods avoid many of the assumptions of model based analysis. Currently, the model based approach dominates experimental work in political science. When analyst both analyze and adjust experimental data, regression models of some form are typically used.

To demonstrate how widespread the model based approach is, we conducted a JSTOR search for articles with experiments using convenience samples published between 1995 and 2007 in the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics*. In our search, we found 47 articles that conducted this type of experiment.² Of these articles, 68% used some form of model based analysis or adjustment with multiple regression and logistic regression being the most commonly used models. As

²We plan on conducting a wider search for papers based on other forms of experimental data for a future version of the paper.

we discuss below, these model based approaches often produce estimates of causal effects that may be inefficient or inconsistent (Freedman 2008b,a,c). We, next, outline the model based approach along with two other methods for adjusting experimental data.

The standard model based method for adjusting experimental data are regression models. Under this strategy, the study groups are given a battery of survey items to measure various covariates that might be imbalanced. Standard items in political science that are measured are gender, party identification, ideology, race, etc. These items are then used post-treatment in a regression adjustment design. Here the outcome from the study is regressed on the treatment measure along with the items that were measured in the experimental session. This method is often referred to as an analysis of covariance or ANCOVA particularly in the psychometric literature. We, next, layout the ANCOVA framework formally.

Let \mathbf{X} be a $n \times k$ matrix of observed characteristics for which we suspect there may not be balance across the treatment and control groups or may be orthogonal to treatment but related to Y_i . This implies that we alter the estimation of the treatment effect as follows

$$\Delta = E(Y_i|\mathbf{X}, D_i = 1) - E(Y_i|\mathbf{X}, D_i = 0). \quad (12)$$

We assume that treatment and outcome are conditionally independent once we condition on \mathbf{X} . If treatment and outcome are conditionally independent based on \mathbf{X} , the estimate of the treatment effect remains unbiased. If we assume that $E[Y_i|\mathbf{X}, D_i]$ is linear then it would seem that a natural estimator for the treatment effect is least squares. More specifically, the analyst might estimate the following linear model

$$Y_i = \beta_0 + \beta_1 d_i + \beta \mathbf{X}_i + e_i \quad (13)$$

In Equation 13, d_i is an indicator for treatment and \mathbf{X} is a matrix of baseline covariates that we think ought to be adjusted for to achieve conditional independence. This assumption of conditional independence is often to as the unconfoundedness assumption (Imbens 2005) or the ignorability assumption (Rubin 1978). The strategy of using the multiple regression model or ANCOVA for statistical adjustment of a randomized experiment is widely used in political science.

While ANCOVA sees widespread use, it has been strongly criticized recently in the statistics literature. Rubin (2005) argues that Fisher’s strong advocacy for ANCOVA is one of Fisher’s greatest mistakes. He demonstrates with a simple example that often ANCOVA does not estimate a causal effect of any kind since force conditioning on a single possible imbalanced covariate can lead to a nonignorable treatment assignment. Freedman (2008b,a) takes up the critique of ANCOVA in even stronger terms. He demonstrates that for the estimation of treatment effects, the multiple regression estimator is biased. The bias goes to zero as the sample size increases, but samples of greater than 500 are needed to reduce the bias to acceptable levels. Worse, asymptotically, estimates from the multiple regression model may be worse than those from a bivariate regression. This is not, however, the greatest deficiency when multiple regression models used for adjustment. The estimated standard errors from the model are inconsistent. The multiple regression model may either overstate or understate the precision by surprisingly large amounts. This is true with both single and multiple treatment regimes. See Freedman (2008b,a) for details. Other model based analyses of experimental data are also problematic. For example, Freedman (2008c) proves that logistic regression models inconsistently estimate treatment effects.

Moreover, while randomization can help produce consistent estimates of treatment effects, it does not imply that any statistical adjustments should be linear and additive. To elaborate, the usual ANCOVA model is

$$Y_i = \beta_0 + \beta_1 d_i + \beta_2 X_1 + \beta_3 X_2 + e_i. \tag{14}$$

While randomization implies an unbiased estimate of β_1 when the model is bivariate, it does not imply that the adjustment for X_1 and X_2 has a linear and additive functional form. The regression model is used due to its ingrained nature, not because there is any aspect of the experiment which suggests the functional form for the adjustment model should be linear and additive.

Rosenbaum (2002a) suggests one simple alternative to ANCOVA. Here covariates that are thought to be related to Y_i are regressed on Y_i . The analyst then applies the standard treatment effects estimator to the residuals from this model. One advantage to this approach is that one can use semiparametric and or robust regression models to overcome the strong

functional form assumptions needed for least squares. One drawback is that the outcome is no longer in an easily interpreted metric. Moreover, Rosenbaum explicitly acknowledges that this method is designed to increase the precision of the treatment effect estimate and does not correct for imbalances. We argue that matching is a better alternative for adjusting experimental research data. We start with an overview of matching.

3.3 Matching

Matching is most closely associated with observational studies, and is generally viewed as a technique that is unnecessary for the analysis of experimental data. We argue that matching provides a useful framework for the statistical adjustment of experimental. We provide a brief overview of matching, but leave a longer discussion of it to others. See Rubin (2006) and Rosenbaum (2002b) for detailed statistical treatments of matching and Sekhon (2008) for an overview in political science.

With observational data, we cannot control selection into treatment and control groups. Here subjects select into the treatment group, therefore the assumption of independence between outcomes and treatment is untenable. As such the counterfactual outcome Y_{i0} is not identified. In a matching analysis, we attempt to identify the counterfactual outcome without randomization. This is possible if we assume that that a set of observed characteristics as measured by \mathbf{X} determine selection into the treatment group. Under this assumption, we may obtain data for a potential control group. This potential control group may not be drawn from the same population as the treatment group, but the observable characteristics in \mathbf{X} match those of the treated units based on a closeness criterion.

The average treatment effect for a matching estimator is

$$\Delta = E(Y_i|\mathbf{X}, D_i = 1) - E(Y_i|\mathbf{X}, D_i = 0). \quad (15)$$

Here the counterfactual outcome is a groups of units that have been matched to the treated along the dimensions of \mathbf{X} . Matching estimators impute the missing value for the control counterfactual using units that are identical or nearly identical to the treated units based on some suitable metric.

The intuition behind matching is fairly simple. A control group is generated by finding subjects that match or nearly match those in the treatment group based on the characteristics in \mathbf{X} . The actual mechanics of matching are more complex. See Sekhon (2008) for a review. Importantly in a matching analysis, observations that do not have matches are discarded. This is due to the assumption of common support between the treated and control units. That is we must assume that for every covariate in \mathbf{X} there is a positive probability of nonparticipation. It may seem strange to drop observations, but if there is no match for a unit, the treatment effect for that unit is not identified. Dropping observations that do not have common support is important since it reduces unobserved bias.

Clearly, matching could be used for post-hoc adjustment of experiments. Here, the analyst would simply match each subject from the control group to a subject in the treatment group. The analyst would then estimate the treatment effect across the matched pairs. Why might we want to use matching as a method for statistical adjustment of randomized experiments with covariate imbalances? First, matching is a fully nonparametric form of adjustment Imbens (2005). When adjusting experiments with multiple regression models, the analyst must assume the adjustment is linear and additive. This is a strong functional form assumption. Adjustment via matching is fully nonparametric, and therefore no assumption about the functional form of the adjustment is necessary. This prevents bias due to using an incorrect functional form for the statistical model. Moreover, adjustment with multiple regression model can obscure a lack of common support across treatment and control groups. With matching, a lack of common support can be easily observed when testing for balance. Finally, if matching is successful, the justification for the assumption of unit exchangeability is strengthened, and the analyst can use randomization tests to estimate treatment effects and draw conclusions about hypothesized treatment effects. This avoids the inconsistent variance estimates from regression models. As Freedman (2008b) points out, there is nothing in the linear regression model that assumes randomization. With matching, one can statistically adjust for covariate imbalance and apply a nonparametric inferential framework that uses randomization as the basis for inferences. See Keele (2008) for a detailed treatment of randomization tests.

Statistical adjustment with matching might take two forms with experiments. The first

form we have already discussed. First, as we discussed previously, one might use design based method that incorporates matching. Here, subjects would be pre-tested for a variety of characteristics including the outcome of interest. Matching would proceed based on the covariates measured in the pretest. This would produce a set of matched pairs, and the analyst would randomize treatment within the pairs. As we discussed previously, this can be viewed as a form of blocking. Such design based matching is advocated by Greevy et al. (2004). They demonstrate that design based matching provides considerable gains in efficiency. Standard matching algorithms, however, cannot be used for this form of matching. Standard algorithms require classification of the treatment and control group. Matching before treatment requires a search to create sets of pairs. Currently, we know of only one publicly available algorithm for this purpose implemented in the `blockTools` library in R. Unfortunately, the algorithm uses greedy matching, which can produce poor matches for some observations due to the sequential nature of the matching. Greevy et al. (2004) use an optimal nonbipartite matching algorithm which does not suffer from this problem, but nonbipartite matching algorithms are not publicly available at this time.

One might imagine an alternative way to proceed. Instead of building matching into the design, the experiment is conducted and during the study a variety of descriptive measures are collected for each subject. Once the experiment is complete, subjects are matched across the treatment and control groups before estimating the treatment effect. Here, matching is an analytic as opposed to design based strategy. Unlike model based strategies, matching requires much weaker assumption and imposed no functional form on the adjustment. One could match on both covariates that are imbalanced to reduce bias and related covariates to increase efficiency. Though if the unobservables are unbalanced matching on observed covariates may be of little use.

Software for matching after treatment is readily available. Here, the analysis proceeds as it would for the analysis of observational via matching. Logically, one-to-one matching would be preserve the experimental structure of the data. In general, we expect greater efficiency gains from pre-treatment matching. The ability to match on the outcome should produce better matches that with post-treatment matching. The drawback to pre-treatment matching is that for many experimental research designs in the social sciences this method is impractical.

Here, analysts must gather data on their subjects and match before conducting the study. This also requires two sessions with subjects which might hinder compliance. Post-treatment matching is undoubtedly easier to implement. It is an open question about whether pre versus post-treatment matching differ in terms of efficiency gains. Post-matching, however, should be superior to ANCOVA. It allows us to avoid a model based design, since matching is a fully nonparametric approach. Moreover, matching allows us to also interpret the resulting estimates as causal parameters, which may not be the case with ANCOVA. Finally, matching allows analysts to easily implement either Fisherian or Neyman inferential techniques. In what follows we conduct a comparative case study of methods for adjustment with experimental data. We compare standard ANCOVA and Rosenbaum's method of covariance adjustment to both pre- and post-treatment matched analyses and unadjusted estimates.

4 A Comparative Method Experimental Design

In our experimental design, we built in numerous methodological comparisons. This will allow us to assess both design and model based adjustments for the resulting data. First, we describe the basic experimental manipulation which did not vary across any of the methodological manipulations. That is all subjects were exposed to the same basic manipulation. The manipulation is taken from the extant literature on priming, and the subjects were randomly assigned to two conditions. In the experiment, we used a manipulation that is designed to activate subjects attitudes toward crime, personal safety, and racial stereotypes. Subjects that were randomly assigned to the treatment group read a fabricated news story on a local mugging. In the story, the race of the assailant and victims was not mentioned. The story appeared to originate in the local newspaper. Subject in the control condition were shown a story from the local newspaper about the iPhone competition with Blackberry. We expect those in the treatment condition to display an increased concern about crime and personal safety. Given the racial priming literature, we also expect exposure to the treatment to activate racial stereotypes. That is we expect treated subjects to rate African-Americans worse relative to whites. To measure both outcomes, we used a series of survey item to construct scales of crime and racial stereotypes. See the appendix for details on the items

that we used to form these scales. We also included a number of survey items that were unrelated to any part of the experiment to disguise our interest in attitudes about race and crime. This experimental manipulation, however, was simply a vehicle for comparing design versus model based adjustments. We now describe the methodological manipulations that were built into the basic experiment.

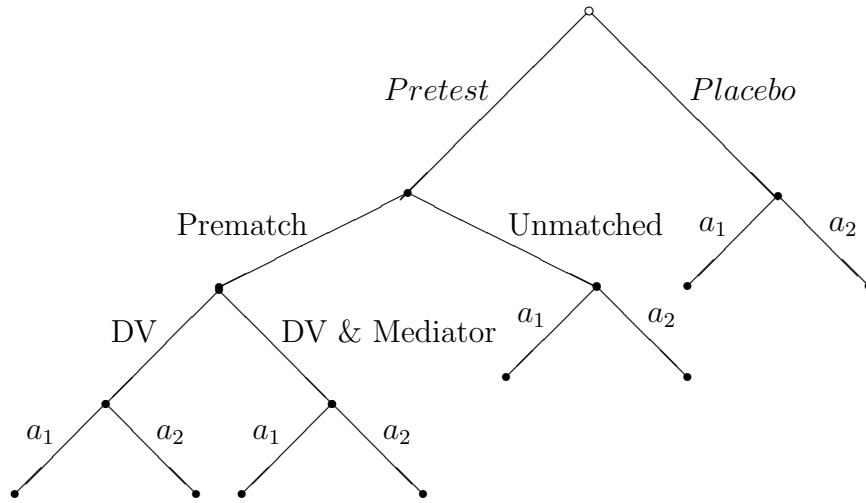
In the experiment, subjects were first asked to attend a session where they completed a prescreening questionnaire. They were then asked to return a week later to participate in the experimental session and complete a post-treatment questionnaire. Subjects were given \$5 on completion of the second session. We offered the financial incentive to decrease attrition among the subjects assigned to the matched pair arm of the study. We discuss attrition rates among the matched pairs subjects in a subsequent section.

First, we investigated the possibility of contamination from a prescreening questionnaire. It is possible that by asking subjects about race and crime prior to the experiment that we activated these attitudes. One fourth of the subjects were randomly assigned to complete a placebo pretest questionnaire. This placebo questionnaire did not contain any items about racial attitudes or crime. In the later analysis, we test whether asking subjects about race or crime pretreatment affected the experimental manipulation. Figure 1 contains a diagram of the full design. Pretesting versus the placebo questionnaire form the first arm of the study. Among the subjects that participated in the pre-treatment questionnaire, half of them then participated in the basic experiment. These subjects were then subsequently analyzed as if we had not collected pre-treatment information. Among these subjects, we used standard model based methods to analyze the data as well as matching on measures collected post-treatment. This forms the second arm of the study in Figure 1.

For the other half of the subjects, we used information from the pre-treatment questionnaire to form matched pairs. We matched the subjects based on several routine background characteristics such as party identification and income among others. Among the subjects that were assigned to be matched pairs, half were also matched on the crime scale, while the other half were matched on both the crime and racial stereotypes scales. This forms the final arm of the design in Figure 1. This design allows us to make three specific comparisons. First, it allows us to assess whether pre-treatment screening can contaminate later parts of

the experiment. Second, it allows us to compare post-treatment analyses based on regression models and matching to pair matching. Third, we can observe the consequences of failing to pre-match on an important predictor of the outcome. We expect the pair matching to be the most powerful design, as it is a design based method of adjustment that takes the outcomes into account when forming the comparison groups, but this method is also costly in the form of the additional time required for the experiment and increases the likelihood of subject attrition.

Figure 1: Matched Experimental Design



4.1 Strategies for Analysis

4.2 Results

We report three different set of results based on our experiment. First, we consider whether pretreatment screening can contaminate subsequent experimental manipulations. Second, we analyze the results from matching the subject into pairs before treatment exposure. Finally, we report results from a comparison of model-based adjustment methods to adjusting the data through matching.

4.2.1 Pretreatment Contaminations

First, we investigate whether exposing subjects to questions on racial attitudes in a pre-treatment questionnaire affected their responses about racial attitudes post-treatment. We analyzed this as a 2 x 2 ANOVA design. That is we compare responses on the racial stereotypes scale across treatment and control and across whether subjects received pre-treatment racial stereotype questions. This allows us to test for two main effects and an interaction effect. First, did the treatment affect racial attitudes? Second, did pre-screening affect racial attitudes? And third was the treatment affect stronger or smaller when subjects participated in the pre-treatment questionnaire? We report only briefly on the first question since it is the focus of later analyses. We found that the main the effect of was highly significant ($p < .01$). The main effect for the pretest versus placebo questionnaire, however, was not significant ($p = .55$).³ This implies that exposing subjects to questions on racial attitudes did not in general affect their post-treatment responses. The evidence for an interaction, however, is mixed. In Figure 2, we report the standard 2 x 2 plot of means. Here, there appears to be evidence of an interaction, in that those who were exposed to pre-treatment racial stereotype survey items had a stronger response to the experimental manipulation. However, we find that this interaction is not statistically insignificant ($p = .406$). While the interaction is not statistically significant, we found that it mattered in subsequent analyses. That is, in later analyses we consistently found that the treatment effect was larger for those exposed to pre-treatment racial attitudes items. As such, we report separate results for those who

³We conducted both a standard ANOVA analysis and a permuted F-test as per the suggestion of Keele, McConnaughey, and White (2008).

received the pre-treatment questionnaire as compared to those who did not.

This suggests some important implications for experimental design. Ideally pretreatment exposures should be unrelated to experimental manipulations, but that is not always the case. There are several design based solutions to this problem. First, the analyst could increase the time between pre-treatment screening and the experimental session. In our case, we waited seven days, so some longer period could be used. Of course, to completely avoid an pre-treatment contamination, pre-treatment screening should be avoided. However, as we will later show, pre-screening and matching can provide large gains in power.

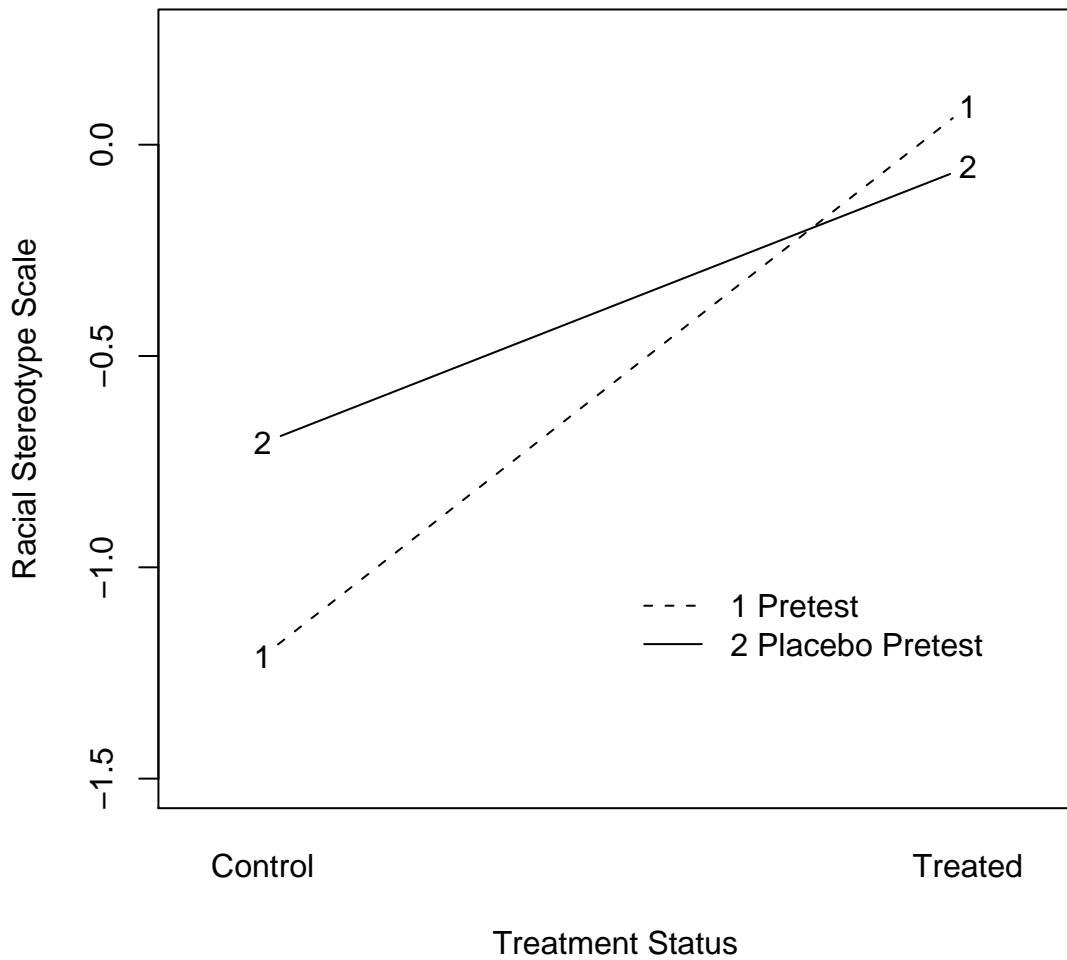


Figure 2: Interaction Between Prescreening Racial Attitudes and No Prescreening

4.2.2 Pair Matching

For half our subjects, we used prescreening information to match the subjects into pairs. We formed two sets of matched pairs. For both sets of matched pairs, we matched on party identification, sex, gender, race, liberal-conservative attitudes, and political knowledge. For half the subjects, we also matched on attitudes toward crime, symbolic racism, and racial stereotypes. For these subjects, the pairs were created based on both the outcome variable as well as the racial attitudes mediator. For the other half of the subjects, we matched on attitudes toward crime in addition to the basic matching covariates. For these matched pairs, we ignored the information provided by matching on the mediation of racial attitudes. We are interested in whether this additional level of equivalency mattered in the formation of the pairs and provided additional gains in power.

To form the matched pairs, we used a greedy matching algorithm with a random starting point for the matches (Moore 2008.). We matched on mahalanobis distances and restricted the range on the matches on the crime and race scales to be within two points. This produced 25 pairs matched on the crime scale and 24 pairs matched on the crime and racial stereotypes scale. We lost four subjects to attrition in each arm of the study. Due to the pairing of the subjects this means that four pairs or eight subjects were lost from each arm of the study reducing the effective number of pairs to 20 and 21. Unfortunately due to a randomization error the number of pairs was further reduced to 10 and 7 in each arm of the study. In general, attrition then was not a severe problem. This may be due to the monetary inducement we used to increase compliance between the pretest screening and the experimental session. To reprise, we expect the treatment to increase scores on the crime scale and decrease scores on the racial stereotypes scale. That is treated subjects should have higher scores on the crime scale than control subjects, and treated subjects should have lower scores on the racial stereotypes scale than control subjects. The results are in Table 3.

For both sets of matched pairs, we find the treatment effects in the expected direction. That is treated subjects were more anxious about crime and rated African-Americans lower in relation to whites in terms of stereotypes. However, the difference was statistically significant for only the pairs matched on crime. The lack of statistical significance for the pairs matched on both crime and racial attitudes may be due to the smaller sample size. Given the small

Table 3: Matched Pairs Results

	Racial Stereotypes	Attitudes Toward Crime	N
Crime Attitudes	-1.49	1.99	
Matched Pairs	0.109	0.009	20
Race and Crime	-1.00	1.11	
Matched Pairs	0.344	0.391	14
Simulation	-1.16	0.60	
Comparison	0.070	0.376	20

Note: First entry is Hodges-Lehmann point estimate. Second entry is exact p -value from sign-rank test.

number of cases, it is remarkable that we find statistically significant treatment effects for any of the groups. Interestingly, the magnitude of the treatment effects here match those found later with much larger sample sizes. Moreover, the effect on attitudes toward crime is highly significant. This suggests that the pair matching does provide a considerable increase in power, but the exact increase in power is not readily apparent. To gain insight into how much power is gained by pair matching, we conducted a simulation exercise. We constructed a comparison group using the subjects that had been assigned to the post-matching condition. To do this, we randomly sampled ten subjects from the treatment group and ten subjects from the control group. With these 20 subjects, we estimated the treatment effect. We then repeated this process sampling a new ten subjects from the treatment and control groups. We iterated this process 1,000 times and calculated the average median shift and exact p -value based on the Wilcoxon sum rank test. These estimates are reported in the last row of Table 3. It appears that the gains in power are considerable for the treatment effect on attitudes on crime, but the gains for the effect on racial stereotypes are negligible. This may be due the fact that we did not match on racial attitudes for this set of matched pairs. This suggests that it is important to match on all outcomes of interest prior to treatment. It would appear then that pair matching does provide considerable gains in power. Analysts must decide whether these gains in power are worth the extra difficulties involved with prescreening and attrition.

4.2.3 Post Hoc Adjustments

While pair matching represents the ideal designed-based form of adjustment, it is not always an option for analysts. Pair matching requires prescreening which may not be possible with many convenience samples. It is often impossible to require multiple sessions with subjects particularly adult subjects. In such situations, analysts can either conduct an analysis without adjusting the data or use some form of post hoc statistical adjustment. The most common form of adjustment is to use a multiple regression model. Here, we compare unadjusted estimates of treatment effects to model based estimates as well to estimates with adjustment via matching. Matching allows for a completely nonparametric form of adjustment avoiding the strong assumptions required for model based adjustments. While unadjusted estimates are equally assumption free, statistical adjustment can correct for accidental imbalances as well as increase the power to detect treatment effects.

In the analyses that follow, we analyze the prescreened data separately from the subjects who received a placebo pretest questionnaire. We find consistently different results across the two groups thus necessitating separate analyses. For the unadjusted estimates, we used the Wilcoxon sum rank test, a permuted t -test, and a linear regression estimator. While these unadjusted estimates should be unbiased, they may not be fully efficient. To increase power, we might adjust for possibly relevant covariates. Covariates that we chose to adjust for are party identification, gender, race, age, liberal-conservative identification, political knowledge, and income. We did find that there were no statistically significant imbalances across these covariates, but adjustment may still improve power. For the model based estimates, we use two forms of ANCOVA. For the first form of ANCOVA, which we call limited ANCOVA, we only adjust for the covariates with largest standardized biases. In the second form of ANCOVA, which we call full ANCOVA, we control for all seven covariates in the multiple regression model. We also used Rosenbaum's method covariate adjustment, where we regressed the outcome variable on all seven covariates. We then applied the Wilcoxon rank sum test to the residuals from this model. Finally, we adjusted the data through matching.

For the matching, we used genetic matching as implemented in the **R** package Matching (Sekhon and Diamond 2005; Sekhon 2007). While for observational data, the analyst must make several decisions about the number of matches and whether to match with replacement.

For experimental data, we must match in a manner that preserves the nature of the experiment. As such, we used one-to-one matches without replacement. We found the genetic matching easily found a balanced set of matches for all the analyses that follow. This isn't surprising given that randomization has done considerable work to make the treated and control groups comparable. For the matched, data, we used the Wilcoxon sign rank test to calculate an exact p -value and the Hodges-Lehmann point estimate for the treatment effect.

We estimate treatment effects for two different outcomes. First, we see whether the manipulation changed the subjects rating of African-Americans relative to whites. Second, we see whether the manipulation changed the subjects feelings about crime and safety. For each outcome, we report two sets of analyses. In the first, we report the point estimate for the treatment effect along with either an exact p -value from a randomization test or an asymptotic p -value for estimation methods that are not exact. Here, we test one-sided hypotheses since we have clear expectations for the direction of the treatment effect. Next, we report on the coverage for each estimation method. Here, we plot the confidence intervals for each estimate to allow for a straightforward comparison of the coverage rates for each method. We, first, report the results of the analysis of the racial stereotypes measure. We also stratify the analysis depending on whether the subject received a placebo pre-test questionnaire or received prescreening questions on race and crime. We stratify the analysis in this way since we find noticeable differences across these two conditions.

Table 4 contains the point estimates and p -values for the seven different estimation methods. First, note that the signs of pretest contamination are obvious in this analysis. For the subjects that did not receive the pretest questionnaire, the effects sizes are less than half those who did. Here, we also find noticeable differences across the model based estimates compared to the unadjusted and matched estimates. In the placebo group, we would conclude that the treatment was ineffective based on the unadjusted estimates, while the model based estimates would lead one to believe the treatment effect was perhaps effective. The matched estimate, however, agrees with the unadjusted estimate but appears to be a bit more precise. We see a similar contradiction among the pretest group. The matched estimate agrees with the unadjusted estimate, while the model-based estimates are have noticeably higher p -values. It is troubling that the model based estimates in both the placebo and

pretest groups tend to disagree with the unadjusted and matched estimates but in different directions. The behavior of the model based estimates here agrees with Freedman (2008a) in that the multiple regression variance estimate can be either too high or too low. Matching, however, either leaves the unadjusted estimate unchanged or improves the precision.

Table 4: Comparison of Estimation Methods For Racial Stereotypes

	Racial Stereotypes Placebo	Racial Stereotypes Pretest
<hr/> Unadjusted Estimates <hr/>		
Rank Sum Test	-0.50	-1.5
	0.109	0.004
T-test	-0.65	-1.30
	0.149	0.015
Bivariate Regression	-0.65	-1.30
	0.127	0.011
<hr/> Model Based Adjustments <hr/>		
Limited ANCOVA	-0.88	-0.95
	0.067	0.057
Full ANCOVA	-0.91	-0.91
	0.067	0.075
Rosenbaum Covariate Adjustment	-0.83	-0.73
	0.056	0.054
<hr/> Matched Adjustment <hr/>		
Sign Rank Test	-1.25	-2.25
	0.143	0.004
<hr/> Note: First entry is Hodges-Lehmann point estimate or mean shift. Second entry is exact or asymptotic p -value <hr/>		

Figure 3 plots the point estimates from each method with 95% confidence intervals. The plot reveals the hazards of relying on the parametric assumptions that undergird regression estimators for treatment effects. As is clear in the plot, the coverage is much tighter for all three regression based estimates. The reason for this is that for the nonregression based estimators negative infinity is contained in the 95% confidence interval. That is without additional parametric assumptions, we cannot estimate a more precise confidence interval. In the experimental setting this is useful information suggesting that additional subjects should be used for greater precision. The regression based estimates simply substitute a parametric assumption to provide tighter confidence intervals. As such, the regression based

estimates are misleading and do not reveal that the confidence intervals are heavily reliant on an untenable parametric assumption. Rosenbaum’s method of covariate adjustment avoids this problem, which further recommends it over standard ANCOVA models.

We next examine the treatment effect estimates for attitudes about crime and safety. Oddly the pre-test contamination appears to work in an opposite direction in this instance. Here, the treatment effect is much larger for the placebo group. The pattern of the effects here is fairly straightforward. For both groups, matching provides noticeable gains in efficiency. The effect is particularly striking for the pretest group. For the unadjusted rank sum test, the exact p -value is 0.27, while for the matched estimate the exact p -value of 0.023, clear evidence that the treatment was effective. There is also further evidence that model-based estimates are unreliable. For the placebo group, the model based estimates tend to disagree with the unadjusted and matched estimate. Moreover, in both the placebo and pretest groups that are noticeable differences between the full and limited ANCOVA models. Here, we see for example the treatment effect fluctuates from 1.68 to with a p -value of 0.07 to 1.53 with a p -value of 0.10. While these differences are not large, they speak to how regression estimates can be highly sensitive to the model specification and lend themselves to arbitrary specification searches for the “best” fit. One might argue that this ability to search for a “best” specification with a regression model is akin to turning experimental data into observational data.

Again, we examine the coverage properties for these different estimators. Figure 4 contains the point estimates and the 95% confidence intervals. The efficiency gains from matching are obvious in the plot. The matching confidence intervals are the only ones that do not contain zero for both groups. This suggests that post-treatment matching is an effective strategy for increasing power to detect treatment effects. Moreover, matching preserves the nonparametric properties of the experiment. It also allows one to easily implement either an estimator for the average treatment effect or a randomization test.

Treatment Effect For Racial Stereotypes

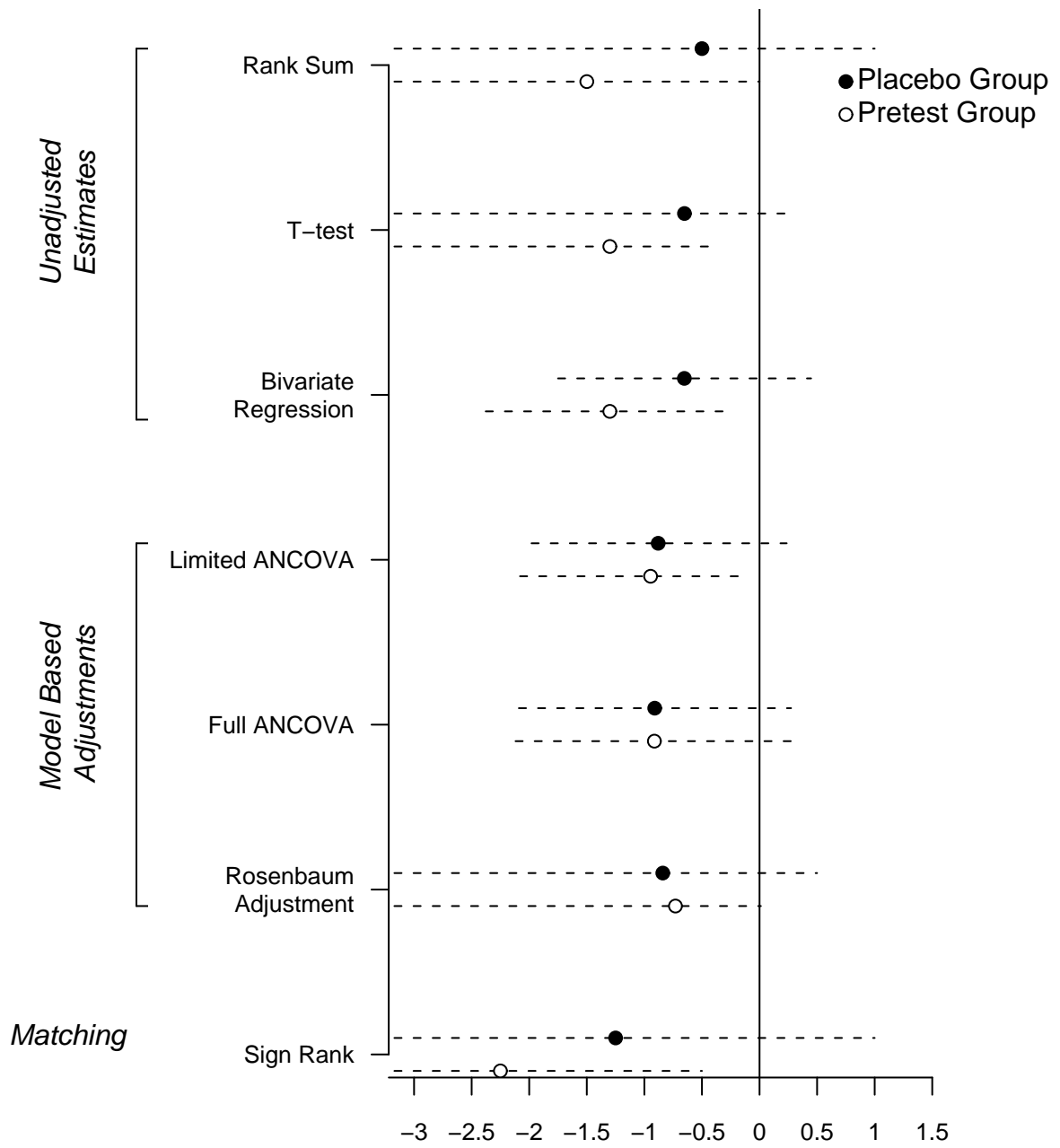


Figure 3: Comparison of Estimates For Racial Stereotypes

Table 5: Comparison of Estimation Methods For Attitudes About Crime

	Crime Attitudes Placebo	Crime Attitudes Pretest
Unadjusted Estimates		
Rank Sum Test	1.5 0.048	0.75 0.268
T-test	1.98 0.052	0.86 0.211
Bivariate Regression	1.97 0.047	0.86 0.206
Model Based Adjustments		
Limited ANCOVA	1.68 0.066	0.99 0.113
Full ANCOVA	1.53 0.097	1.10 0.089
Rosenbaum Covariate Adjustment	1.55 0.068	1.16 0.067
Matched Adjustment		
Sign Rank Test	2.25 0.017	1.75 0.023
Note: First entry is Hodges-Lehmann point estimate or mean shift. Second entry is exact or asymptotic p -value		

Treatment Effect For Crime Attitudes

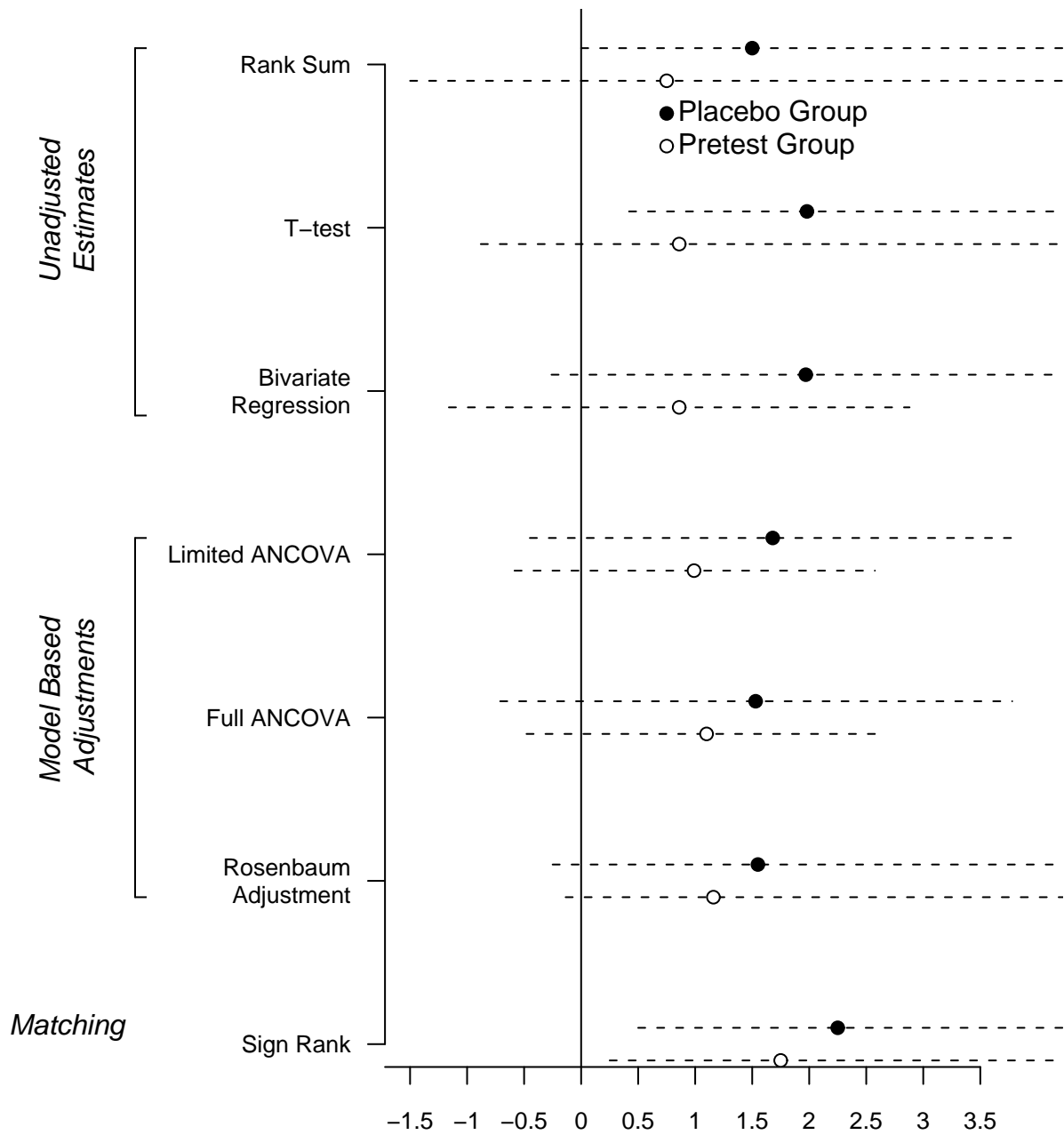


Figure 4: Comparison of Estimates For Attitudes Toward Crime

4.3 Post Hoc Experimental Analysis: A Study of Support for the Iraq War

In our second example, we analyze data from White (2003). In this example, no design-based adjustment were included in the experiment. Here, we explore options for adjusting the data post-treatment only. White designed an experiment to test the effect of a media source cue on support for the Iraq war. All subjects viewed a news magazine article laying out arguments for opposition to the war in Iraq; the article was manipulated across conditions to vary the source of the article. The article was presented inside a news story appearing in either a black news magazine (*Black Enterprise*) or a mainstream news magazine (*Newsweek*). Subjects were then asked to report on a 1-7 scale whether the Bush administration had presented enough evidence to go to war and whether he or she favored or opposed U.S forces in Iraq. The experiment was run separately on both white and black subjects, as the theory suggested that blacks and whites would respond differently to the treatments. It was hypothesized the blacks might be less supportive of the war when presented with the news story in a news source with a clear racial focus. In the analysis that follows, we only report results for black respondents.

The subject pool is a mix of students and adults recruited on and around a Midwestern and Southern university. Here, we expect adjustments to be more powerful given the more heterogenous subject pool than our last example. We checked balance across the treatment and control groups for the following covariates: party identification, liberal-conservative identification, level of political knowledge, an indicator for whether they were related to armed services personnel, sex, whether the respondent was married, level of education, whether the respondent was employed or not, income, and whether they owned their own home. We assessed balance using a randomization based test (Hansen and Bowers 2008). We present the results from the balance tests in Table 6. We see that four covariates were not perfectly balanced by the randomization, and the global balance test indicates a statistically significant level imbalance. In such situations, post hoc adjustment should increase power noticeably. We, again, use one-to-one genetic matching without replacement to match control subjects to treated subjects. Table 6 also includes the balance test results for the matched data. None of the covariates now display statistically significant imbalances and the global test is no longer

statistically significant.

Table 6: Balance Test for Iraq War Experiment

	Standardized Bias Unmatched	Standardized Bias Matched
Party Identification	-0.25	0.00
Liberal - Conservative	-0.38*	0.02
Knowledge	0.37*	0.02
Related to Armed Services Personnel	-0.13	-0.07
Sex	-0.19	0.11
Married	-0.17	0.08
Education	0.06	0.11
Employed	-0.43*	0.28
Income	0.55*	0.00
Own House	-0.15	-0.32
Gobal χ^2 Value	22.77*	8.28
* p -value < 0.05		

We present the results for the seven different estimators in Table 7. Here, we see a very similar pattern of effects across both the unadjusted and model based estimates. One exception is Rosenbaum’s method, which returns treatment effect estimates that are approximately half the other model based estimates. This is due to the fact that the data are unbalanced and this method does not correct for such imbalances. What is most noticeable from the results in Table 7 are the gains in efficiency due to matching. As we would expect given that matching corrected several significant imbalances, the treatment effect is now more precisely estimated. This is readily apparent in Figure 5 which contains plots of the point estimates and 95% confidence intervals. Only the matched estimates are clearly bounded away from zero. As in the last example, matching provides noticeable gains in power without requiring the strong assumptions necessary for model based adjustments. Here, we were able to correct several significant imbalances that occurred despite randomization.

5 Conclusion

While experiments are clearly the “gold standard,” this does not imply that they are without complications. As we have outlined, imbalance, causal heterogeneity, inefficiency, and equiv-

Table 7: Comparison of Estimation Methods For Opposition to Iraq War

	Iraq Outcome 1	Iraq Outcome 2
Unadjusted Estimates		
Rank Sum Test	0.00	0.00
	0.320	0.058
T-test	0.52	0.49
	0.078	0.086
Bivariate Regression	0.52	0.489
	0.078	0.086
Model Based Adjustments		
Limited ANCOVA	0.59	0.71
	0.073	0.028
Full ANCOVA	0.64	0.66
	0.06	0.042
Rosenbaum Covariate Adjustment	0.18	0.37
	0.33	0.103
Matched Adjustment		
Sign Rank Test	1.75	1.49
	0.011	0.000
Note: First entry is Hodges-Lehmann point estimate or mean shift. Second entry is exact or asymptotic p -value		

Treatment Effect For Iraq Support

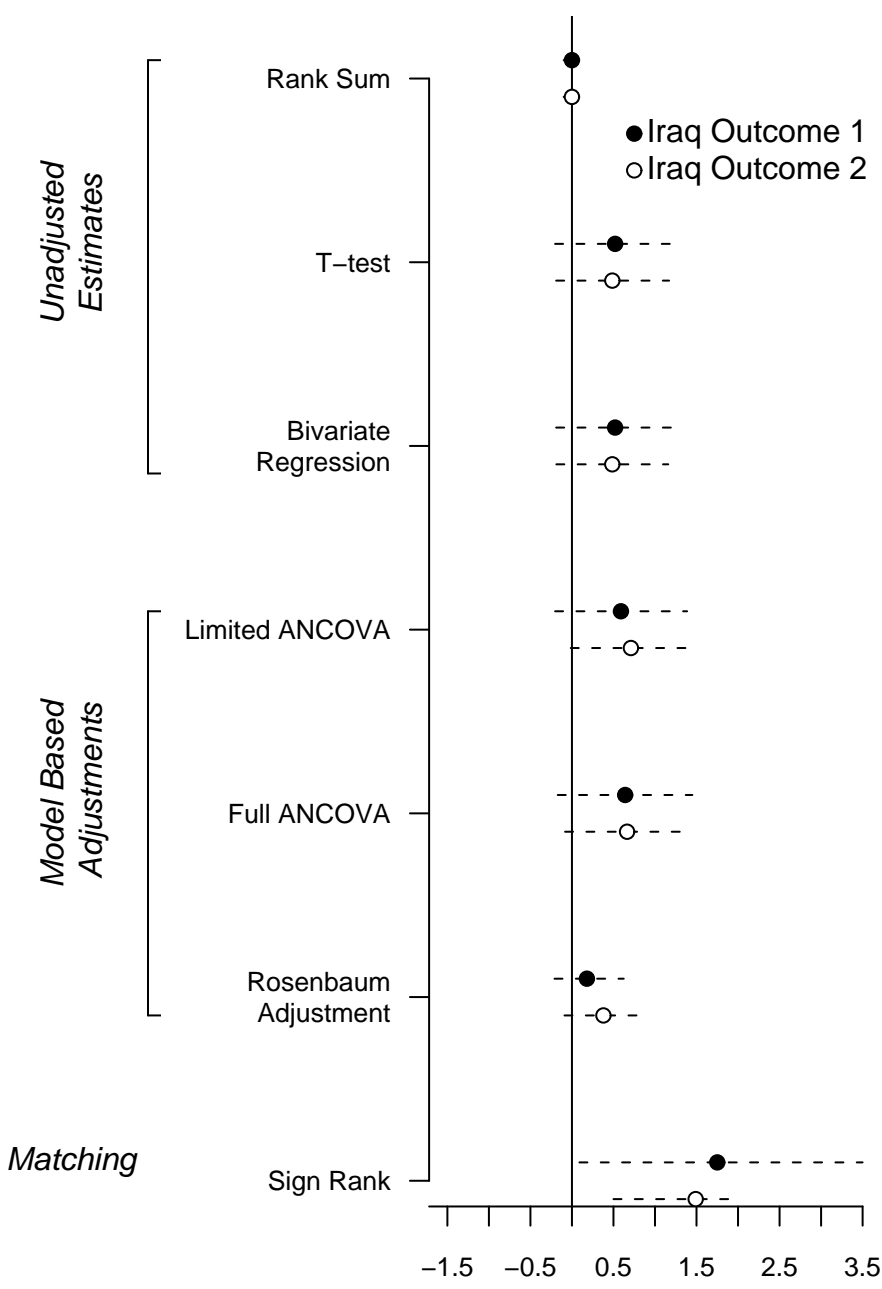


Figure 5: Comparison of Estimates For Belief in Iraq War Arguments

alence violations are all possible in a single experiment. Current practice in political science relies heavily on regression model based approaches to cope with these problems. In truth, these are quite possibly the worst method for adjusting experimental data. Regression based methods have the strongest assumptions that are not required with experiments. We found in our analyses that regression method did not adjust as well as post-treatment matching, moreover the results from ANCOVA in one instance differed from both the unadjusted estimates and the matched estimates. This points toward the problems noted in Freedman (2008b).

While the pair matched design appeared to be the most powerful design this was only true when we correctly pre-matched on relevant characteristics. Pair matching did not appear to improve power when we failed to match on racial stereotypes. Pair matching also requires prescreening and may induce additional attrition. In some settings, it may not be possible to pre-screen. When only post-treatment adjustment is possible, standard matching techniques provide an alternative to the usual model based approaches. Matching imposes fewer assumptions on the data and allows the analyst to adjust not only for efficiency but also accidental imbalances. We found in both the Iraq data and in our experiment on priming that adjusting through matching often increased power. While careful planning is required in both the design and analysis of experiments, we feel that some of the techniques presented here are superior to much of current practice.

Appendix

In the prescreening questionnaire and post-test questionnaire, we asked the subjects to respond to a number of survey items. Many of these were routine questions about party identification, sex, race, and so forth. The outcomes of interest in the experiment, however, were racial stereotype ratings and attitudes about crime and personal safety. For these outcomes, we used a set of scales. The specific items that comprised these two scales are below. We used these same two scales in the matching process to form the matched pairs based on pretest information.

Crime and Personal Safety Scale To form a scale of attitudes toward crime and personal safety, we simply summed the response from each item to form an additive scale. The Cronbach's α score for this scale was .72

- On a scale from 1 to 10, how safe would you say the OSU CAMPUS AREA is? Where 1 is very safe and 10 is very dangerous
- On a scale from 1 to 10, how safe do you feel walking alone in the area just outside of the OSU campus after dark? Where 1 is very safe and 10 is very dangerous.
- Thinking about your day to day life in Columbus how concerned are you that you will be mugged on the street
- Thinking about your day to day life in Columbus how concerned are you that you will be a victim of a violent crime
- How concerned do you think OSU students should be about crime?
- How concerned are you that a member of your family or a close friend might one day be a victim of a violent crime.

Racial Stereotypes Scale For each racial stereotype item, we took the difference of the subjects rating of whites and African-Americans and summed these differences to form a racial stereotypes scale. The Cronbach's α score for this scale was .71.

- Generally speaking, how well does the word VIOLENT describe WHITE AMERICANS as a group?

- Generally speaking, how well does the word VIOLENT describe AFRICAN AMERICANS as a group?
- Generally speaking, how well does the word INTELLIGENT describe WHITE AMERICANS as a group?
- Generally speaking, how well does the word INTELLIGENT describe AFRICAN AMERICANS as a group?
- Generally speaking, how well does the word LAZY describe WHITE AMERICANS as a group?
- Generally speaking, how well does the word LAZY describe AFRICAN AMERICANS as a group?

References

- Bowers, Jake, and Ben Hansen. 2007. "Fixing Broken Experiments: How to Bolster the Case for Ignorability with Full Matching." Unpublished Manuscript.
- Fisher, Ronald A. 1935. *The Design of Experiments*. London: Oliver and Boyd.
- Freedman, David A. 2008a. "On Regression Adjustments in Experimental Data." *Advances in Applied Mathematics* Forthcoming.
- Freedman, David A. 2008b. "On Regression Adjustments in Experiments with Several Treatments." *Annals of Applied Statistics* Forthcoming.
- Freedman, David A. 2008c. "Randomization Does Not Justify Logistic Regression." Unpublished Manuscript.
- Greevy, Robert, Bo Lu, Jeffery H. Silber, and Paul Rosenbaum. 2004. "Optimal Multivariate Matching Before Randomization." *Biostatistics* 5 (April): 263-275.
- Hansen, Ben B., and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified, and Clustered Comparative Studies." *Statistical Science* Forthcoming.
- 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 (March): 481-502.
- Imbens, Guido W. 2005. "Nonparametric Estimation of Average Treatment Effects." *Review of Economics & Statistics* 86 (February): 4-29.
- Keele, Luke. 2008. *Semiparametric Regression for the Social Sciences*. Chichester, UK: Wiley and Sons.
- Keele, Luke, Corrine McConnaughy, and Ismail White. 2008. "Randomization Tests With Experimental Data." Working Paper.
- Moore, Ryan T. 2008. "Blocking to Improve Political Science Experiments." Unpublished Manuscript.
- Neyman, Jerzy. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5 (November): 465-472. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).
- Rosenbaum, Paul R. 2002a. "Covariance Adjustment In Randomized Experiments and Observational Studies." *Statistical Science* 17 (August): 286-387.
- Rosenbaum, Paul R. 2002b. *Observational Studies*. 2nd ed. New York, NY: Springer.
- Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6: 34-58.
- Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100 (March): 322-330.
- Rubin, Donald B. 2006. *Matched Sampling For Causal Effects*. New York, NY: Cambridge University Press.

- Sekhon, Jasjeet S. 2007. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package For R." *Journal of Statistical Software* Forthcoming.
- Sekhon, Jasjeet S. 2008. "The Neyman-Rubin Model of Casual Inference and Estimation via Matching Methods." In *The Oxford Handbook of Political Methodology*, ed. Janet Box-Steffensmeir, Henry E. Brady, and David Collier. Oxford Handbooks of Political Science Oxford: Oxford University Press.
- Sekhon, Jasjeet S., and Alexis Diamond. 2005. "Genetic Matching for Estimating Causal Effects." Presented at the Annual Meeting of the Political Methodology, Tallahassee, FL.
- White, Ismail K. 2003. "Racial Perceptions of Support for the Iraq War." Unpublished Data.