

# Adjusting Experimental Data: Models versus Design\*

Luke Keele<sup>†</sup>      Corrine McConnaughy<sup>‡</sup>      Ismail White<sup>§</sup>

September 24, 2010

## Abstract

Randomization in experiments allows researchers to assume that the treatment and control groups are balanced with respect to all characteristics except the treatment. As such, treatment effects can be estimated with a simple difference of means—no other covariates are necessary. Still, analysts often adjust using experimental data for additional covariates with statistical models. In this paper, we address the issue of adjustment with experimental data. First, we outline the motivations for adjustment of experimental. We then review design based and analytic strategies for addressing these underlying motivations. We offer a comparison of the most common method of adjustment—the standard least squares regression to estimate an analysis of covariance (ANCOVA) model—to two methods of statistical adjustment that avoid ANCOVA’s strong functional form assumption: Rosenbaum’s method of covariate adjustment and matching. For all methods of adjustment, we outline a design and analytic strategy that maximizes transparency and illuminates whether inferences are adjustment dependent. Finally, we evaluate these methods of adjustment and demonstrate our design strategy with an original experiment on racial priming.

---

\*Authors are in alphabetical order. We thank Jas Sekhon, Rocio Titunik, Kosuke Imai, Don Green and Ben Hansen for comments and discussion. We thank Kevin Duska for research assistance. A previous version of this paper was presented at the 2008 Annual Meeting of the Society of Political Methodology, Ann Arbor, MI, the 2010 Annual Meeting of the Midwest Political Science Association, and the 2010 Annual Meeting of the American Political Science Association, Washington D.C.

<sup>†</sup>Associate Professor, Department of Political Science, 2137 Derby Hall, Ohio State University, Columbus, OH 43210 Phone: 614-247-4256, Email: keele.4@polisci.osu.edu

<sup>‡</sup>Associate Professor, Department of Political Science, 2018 Derby Hall, Ohio State University, Columbus, OH 43210 Phone: 614-292-9658, Email: mcconnaughy.3@polisci.osu.edu

<sup>§</sup>Associate Professor, Department of Political Science, 2008 Derby Hall, Ohio State University, Columbus, OH 43210 Phone: 614-292-4478, Email: white.697@osu.edu

Researchers looking to estimate causal effects have come to regard the randomized experiment as “the gold standard.” The gold standard label results from the ability of randomization to produce equivalence across 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 pretreatment levels on the dependent variable and their propensity to respond to the treatment.

Randomization, of course, does not obviate the need for statistical analysis of experimental data; however, depending on the design this can be a relatively simple comparison of means across treatment conditions. The current dominant practice in political science, however, is to analyze experimental data with regression models of some type. More often than not, multivariate regression models are used so that the analyst can statistically “adjust” the experimental data for other measured covariates. Whether this strategy is actually serving the inferential needs of the investigator, and whether it is the best available option, however, is too often an unanswered question.

In this paper, we take up this question by focusing on what may appear to be two disparate topics: statistical adjustment of experimental data and experimental design. Clearly, experimental investigators understand the need to think critically about experimental design. The focus of design, however, tends to be on developing a critical test of the theory. Design, then, is understood as a separate process from statistical analysis, which occurs once the study is complete. Thus, decisions about design and analysis are often separated from each other temporally by the actual execution of the experiment itself. Often the design of the experiment itself, however, can eliminate the need for statistical adjustments. We explore how design can intersect with the adjustment of experimental data to eliminate the need for statistical adjustment. Yet, we also examine the limitations of design based strategies. When adjustment is necessary, we suggest alternatives to the standard model based methods, which can often lead to inferential errors and a corruption of the randomization in the experiment. Finally,

we outline a basic set of “best practices” for the reporting of experimental results that allows readers to understand the role of adjustment in their inferences.

# 1 Assumptions, Experiments and Statistical Analysis

Before exploring adjustment strategies—design or model based—we outline an approach to the analysis of experiments. This approach need not be confined to experiments since it is simply a reasoned argument about the role of assumptions in statistical analysis, but here we develop the argument within the context of experiments. The analysis of data invariably requires the use of assumptions. The question is not whether assumptions must be made, but is instead about the quality of the assumptions. Statistical assumptions can be divided into two types: refutable and nonrefutable. Refutable assumptions can be evaluated with observed data while nonrefutable assumptions cannot be tested with any configuration of the data Manski (2007). With observational data, nonrefutable assumptions are unavoidable. Next, we discuss this distinction between assumptions in the context of experiments.

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. Here, it is useful to define some notation. 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. The observed outcome is determined by a 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 written as  $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$ . The individual level treatment effect for unit  $i$  is  $\Delta_i = E[Y_{i1}] - E[Y_{i0}]$ . For this to be

true we must make the following assumption:

$$Y_{i1}, Y_{i0} \perp T_i \tag{1}$$

or in words the potential outcomes must be independent of the treatment. Randomization of the treatment,  $T_i$  makes this true in expectation. Why? When treatment is randomized, we know the following to be true:  $Pr(T = 1) = Pr(T = 0)$ . With observational data, a related but different assumption of the following form

$$Y(1), Y(0) \perp T_i \mid \mathbf{X}. \tag{2}$$

must be made. This assumption is often referred to as conditional ignorability since ignorability is now conditional on the observed covariates  $\mathbf{X}$ . This ignorability assumption was first articulated by Rosenbaum and Rubin (1983) who referred to it as “ignorable treatment assignment.” Others have referred to it as “selection on observables” (Barnow, Cain, and Goldberger 1980). Under this assumption, the investigator asserts that all variables that need to be adjusted for are observed. Often in econometrics texts, unbiasedness of least squares estimators are appealed to on the grounds that the model is “correctly” specified. This is simply a tacit ignorability assumption. In a randomized experiment we don’t have to assume that Equation 1 is true since the physical act of randomization will guarantee it to be true within some random error as the sample size grows. Without randomization, causal inference must proceed by asserting that either Equations 1 or 2 are true by assumption. And unless randomization has occurred, either assumption is nonrefutable. In short, we have to assume that there are no unobserved confounders. In an experiment, this assumption is replaced by a mechanism, randomization over which the researcher has complete control.<sup>1</sup>

In the context of a randomized experiment, we must also assume there is no “interference” between different units. That is the observation on one unit is unaffected by

---

<sup>1</sup>Of course, while the researcher has complete control over randomly assigning units implementation of that randomization may confront any number of compliance issues, which are beyond the scope of this article.

the particular assignment of treatments to the other units. We must also assume that there do not exist any unrepresented versions of the treatment. These two assumptions are commonly combined under the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1978). Under this assumption, a simple, consistent estimator for the average treatment effect (ATE) is the observed difference between the control group and the treatment group in their average posttreatment values on the outcome variable  $Y_i$ . If  $m$  of  $n$  subjects in the experiment received the treatment, then the estimator is defined as:

$$\begin{aligned}\hat{\Delta} &= \frac{1}{m} \sum_{i|T_i=1} Y_{i1} - \frac{1}{n-m} \sum_{i|T_i=0} Y_{i0} \\ &= \bar{y}_1 - \bar{y}_0\end{aligned}\tag{3}$$

Moreover, most analysts also assume large sample theory holds and that the test statistic follows a  $t$ -distribution. This may not be true if the distribution of the data has heavy tails or outliers. With experiments, however, we need not make this parametric assumption. Fisher (1935) demonstrated how randomization provides a “reasoned basis” for inference that avoids parametric assumptions. Thus analysts can instead use a series of nonparametric tests known as randomization test (Keele, McConnaughy, and White 2008). Let’s say an analyst makes the large sample assumption and concludes the treatment is without effect. If this assumption had not been made and randomization inference had been used the analyst would have correctly concluded that the treatment was effective. In this example, one could say that the power of randomization has been wasted by the poor use of assumptions. Moreover, the effort that went into developing a design that matches the theory has been wasted as well.

Thus it would seem that a point worth repeating is that assumptions need to be made with care even with experiments where randomization obviates the need to assert ignorability. The fact that randomization has occurred gives the investigator considerable power, and it would be tragic to waste that power. Therefore, one insight that might be gleaned from this exercise is that one should make the fewest number of

reasonable assumptions in the analysis of experimental data. Otherwise, unwarranted assumptions may undo the leverage provided by randomization. We argue that using statistical models to adjust experimental data often adds unnecessary assumptions that can undo what is gained by randomization. We explore how design can help avoid these assumptions and develop an analytic method that clarifies the role of adjustment if used. We now review why one might adjust experimental data with statistical models.

## 2 Adjusting Experimental Data

As we show shortly, it has become widespread practice to adjust experimental data through the use of multivariate regression models. Here, we explore the logic behind why one might want to make such adjustments. We begin with an overview of the estimation of treatment effects with regression models. In the following regression model

$$Y_i = \beta_0 + \beta_1 T_i \tag{4}$$

$Y_i$  is the observed outcome for the units with  $i = 1, \dots, N$ . Here  $T_i$  indicates treatment,  $T = 1$  for treated,  $Z = 0$  for control. In what follows, we assume  $T_i$  has been randomly assigned to units and this is a random vector. The least squares estimate,  $\hat{\beta}_1$ , is the estimate of  $\Delta$ , the treatment effect, in this model. Formally, the least squares estimate of  $\Delta$  is

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

where  $\tilde{p}$  is

$$\bar{T} = \frac{1}{n} \sum_{i=1}^n T_i = \tilde{p} \tag{6}$$

or the observed proportion of the sample that receives the treatment.

Why might we introduce additional covariates into this regression model? There are two reasons. First, randomization will balance observed and unobserved covariates in expectation. Within a single experiment, a covariate may not be balanced by the

randomization. Let's review the logic for using a regression model when randomization does not fully balance the data. Assume there is an additional pretreatment covariate or a set of covariates,  $X_i$ , that may affect the level of  $Y_i$ . We 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:

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

We find that the expectation of this estimate is

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

If accidental covariate imbalance occurs then  $Cov(T_i, X_i) \neq 0$  and if  $\beta_2 \neq 0$ , this results in a biased estimate of the treatment effect  $\beta_1$ . For the derivation below, we assume a constant treatment effect, which implies that  $\beta_1$  does not vary with each unit in the analysis. See Freedman (2008b,a) for derivations without this assumption. We can write the bias in the estimate of the treatment effect as

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

If we take the probability limit of equation 8, it follows that as  $n \rightarrow \infty$  then  $\bar{T} \rightarrow p$ , where  $p$  is the limit of  $\tilde{p}$  as  $n$  increases. Now, we can rewrite this equation as:

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

Asymptotically the bias is:

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

Therefore, the bias due to an unbalanced covariate is simply the difference in means across the treatment and control group on the unbalanced covariate. This derivation reveals the logic for using regression models when imbalance is present after randomization. As such, many investigators use a multivariate regression model to adjust for accidental imbalance in a covariate. See Bowers (2010) for a full discussion of how regression adjusts for additional covariates.

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. In this instance, including  $X_i$  in the estimating equation may cause the following to be true:  $\sigma_{\hat{\beta}}^2 < \sigma_{\beta}^2$ . Including  $X_i$  on the right hand side of the model may 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. Thus another reason to use a multivariate regression model is to increase the precision in the estimate of the treatment effect. Using multivariate regression models with experimental data is often referred to as an analysis of covariance or ANCOVA particularly in the psychology literature.

In short, there are two different reasons for why analysts might adjust experimental data: increasing precision and removing bias due to imbalance.<sup>2</sup>

There is, however, a fundamental asymmetry between these two reasons for adjustment. When the experiment produces balance in pretreatment covariates, this gives researchers evidence that the randomization procedure was valid and it was implemented without errors or deviations from protocol. In this scenario, researchers have no reason to suspect that there is hidden bias. Here, covariate adjustment can be done to increase precision, but this should not affect the point estimates in any significant way. When the experiment fails to produce balance in pretreatment covariates, the situation is very different. Of course, in any given realization a valid randomization procedure can yield a highly imbalanced treatment and control group. When this occurs, however, the researcher has no way of distinguishing between whether he or she

---

<sup>2</sup>In truth, one might also use treatment-covariates interactions to model treatment effect heterogeneity in a regression model as well. We do not consider this form of adjustment.

is just unlucky or whether the randomization procedure been implemented incorrectly. One would think that the former would be better than the latter, but the distinction is not relevant if the imbalance is severe. If pretreatment covariates that are thought to be correlated with the outcome are severely imbalanced, then we can have little confidence in our results because now there is no reason to believe that there is no imbalance in unobservables. Once pretreatment covariates are severely imbalanced, an experiment reverts to being an observational study. We can do covariate adjustment and present the results, but the independence between  $Y$  and  $T$  conditional on  $X$  now has to be assumed, because we have no confidence in the randomization. Therefore when there is imbalance and especially when the imbalance is severe, the problems that adjustment seeks to solve are much more serious than when the adjustment is used to increase precision, because in the latter case we have lost confidence in the randomization. In the latter case, we are asking much more from the adjustment than we are asking in the former, and whether the adjustment succeeds in achieving the goal of yielding unbiased inferences is ultimately unobserved.

We now review the use of regression models to adjust experimental data in political science where we argue the asymmetry behind the logic for adjustment has largely been lost. We conducted a review of the analysis of experiments in three journals: *American Political Science Review*, *American Journal of Political Science*, and the *Journal of Politics* from 1995 to 2008. We then calculated the percentage of experiments where a multivariate regression model of some type was used with the experimental data. In the *APSR*, we find that 95% of the experiments were analyzed with either least squares or binary choice model such as logistic regression. The percentages were 95% and 74% in *JOP* and *AJPS* respectively. Analysts often report unadjusted estimates as well this is not standard. For example in the *APSR*, 13% of the articles did not present unadjusted estimates. Rates were nearly identical in *JOP* and *AJPS*. Clearly most analysts choose to adjust their data with multivariate statistical models. We next explore why the use of statistical models is not always advisable.

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 possibly imbalanced covariate can lead to nonignorable treatment assignment. Freedman (2008*b,a*) takes up the critique of regression models 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. The bias arises from the fact that the linear model assumes treatment effects are constant across units (Freedman 2008*b,a*). Worse, asymptotically, estimates from the multiple regression model may be worse than those from a bivariate regression. This is not, however, the greatest deficiency of multiple regression models used with experimental data. In these models, 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. Why should this be the case?

Most textbook derivations of least squares properties assume fixed or nonstochastic regressors, though derivations with stochastic regressors are common enough (Greene 2000). These derivations, however, do not hew closely to what occurs in an experiment. When randomization occurs, both the outcome and error terms are fixed with only the treatment assignment vector being stochastic. This implies that the random properties of the error term are due entirely to the fact that treatment is random. Of course, if this is true the error term is now strongly dependent on treatment assignment, and of course is no longer independent and identically distributed (IID) across the study units. The usual Gauss-Markov assumptions for the linear regression model, of course, hold that the error terms are IID. The difficulty is that this assumption is directly contradicted by randomization: the errors will vary with treatment by definition making the error variance nonconstant. Of course, nonconstant error variance in a regression model is usually referred to as heteroksedasticity. See Freedman (2008*b,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 for binary outcomes.

Moreover, randomization does not imply that the nature of statistical adjustments should be linear and additive. To elaborate, the usual ANCOVA model is

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_1 + \beta_3 X_2 + e_i. \quad (11)$$

While randomization implies an unbiased estimate of  $\beta_1$  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. We acknowledge that using models to adjust experiments, despite the concerns outlined above, is often unproblematic. Our point is that models with experimental data require additional assumptions that may not be warranted.

Regression models can also lead to fishing expeditions for the “best” set of covariates to adjust for. That is, researchers may try different sets of control variables until the experiment “works.” Thus researchers may find one particular specification that showcases the finding from the study. An experiment then becomes as open to the type specification searches that are common with observational data. It is hard to know how often this occurs, but it is certainly possible. Indeed, it is specification searches of this sort that have caused the FDA to require that all covariates used to adjust experimental results must be declared in the protocols before the execution of drug trials. Researchers then are only allowed to adjust for covariates declared in the pretreatment protocols.

## 3 Alternatives to Model Based Adjustments

In this section we explore alternatives to regression model based adjustments. One of these methods is fully design based—the design of the experiment addresses the motivating reasons for statistical adjustment. The other two methods are not design based, but avoid some of the concerns that arise when regression is used for adjustment.

### 3.1 Blocking

Blocking is a design based strategy for increasing both balance and the precision of the treatment effect estimate. In a traditional block design, the analyst stratifies the randomization along levels of some possibly confounding or variance inflating factor(s). If the investigator is concerned about an imbalance on race, for example, randomization would be implemented within stratified groups or “blocks” of racial categories. Some authors have recently called for greater usage of block designs in experimental research (Imai, King, and Stuart 2008). They demonstrate that the use of blocking can never hurt the estimate in terms of power. That is blocking will never decrease the precision of the treatment effect estimate even if blocking is unnecessary (Imai, King, and Stuart 2008). Under a traditional block design, it is usually difficult to block on more than one factor unless the number of experimental units is large.

New Matching algorithms provide significant improvements over traditional block designs. The analyst can use a matching algorithm to form blocks based on a multi-variate measure of units’ (subjects’) distance on a large number of covariates (Moore 2008). For example, an analyst might use these algorithms in a matched pair design. The matched pair design is simply a block design where each block contains only two units. In a matched pair design, experimental units are placed into pairs and the randomization is performed within the matched pairs. Typically subjects are put into pairs based on a single covariate. Search algorithms make this matching process much less ad hoc and allow for creating matched pairs based on a number of covariates. Moreover, the software tools developed by Moore (2008) can create matched sets of 3

or more. Thus the randomization can occur within matched groups larger than pairs.<sup>3</sup> Greevy et al. (2004) study the effects of matching subjects before treatment. They demonstrate that blocking via matching provides considerable gains in efficiency. Imai (2008) proves that there is almost always gains in efficiency for pair matched experiments compared to standard experiments. Block designs based on matching greatly reduce the likelihood of accidental imbalance on any covariate used to create the blocks. The advantage to this method is that the adjustment is built into the design of the experiment itself. Here, the logic and role of adjustment is completely transparent. We might say that if experiments are the gold standard for the estimation of causal effects, adjustment via matched-based block designs is the gold standard of adjustment.

The drawback to block designs, however, is that they require prescreening of the covariate(s) that might cause imbalance concerns. That is blocking almost always requires additional contact with the subjects: one for prescreening and then a second for the actual experiment. With two sessions, questions of compliance can arise. That is subjects may not return to the second session without inducement. Contamination concerns arise when measurement of the covariates of concern interferes with subjects' response to the treatment, changing the observed levels of  $Y_i$ . Again the design can help alleviate possible contamination. First, one can use a longer time interval between screening and the experimental session. Second, one can randomly allocate the some subjects to receive sensitive question items while others do not. This allows for a complete understanding of any possible contamination. Thus, the objections to design-based adjustment are practical and not statistical.

---

<sup>3</sup>It is important to note that standard matching algorithms are inappropriate for this purpose. Standard algorithms require the analyst to classify subjects into treatment and control groups, but matching before treatment requires a search to create sets of pairs or groups. 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 (Rosenbaum 1989). 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.

## 3.2 Alternatives to Model-Based Adjustments

If a block design is not used, any adjustment must be done after execution of the experiment. As we outlined above, the usual choice is a regression model of some sort. Next, we outline some alternatives to regression models. The first method we outline is only designed to increase the precision of the treatment effect estimate and does not correct imbalance. The second method, matching, is nonparametric but creates new complications.

### 3.2.1 Rosenbaum's Method of Covariate Adjustment

Rosenbaum (2002a) suggests one simple alternative to ANCOVA. 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. We should note that Rosenbaum explicitly acknowledges that this method is designed to increase the precision of the treatment effect estimate and does not correct for accidental imbalances. This strategy is not strictly speaking design based, but it does make data mining less likely. Here, since estimation of the regression model is separated from treatment effect estimation, it is less likely that analysts will perform specification searches. One primary advantage of this method of adjustment, is that one can simply conduct a test of the sharp null while incorporating covariates.

### 3.2.2 Matching

Matching is most closely associated with observational studies and is not widely applied to experimental data. We argue that matching provides a useful framework for the statistical adjustment of experimental data as well. 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. See Bowers (2010) for another perspective on the use of

matching with experimental data.

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  $\mathbf{X}$  a matrix of pretreatment covariates. Here the counterfactual outcome is a groups of units that have been matched to the treated along the dimensions of  $\mathbf{X}$ . Clearly, matching could also be used for post-hoc adjustments to experimental data. Instead of building matching into the design, the experiment is conducted and during the study (before treatment) a variety of descriptive measures are collected for each subject. Once the experiment is complete, subjects are matched on pretreatment covariates before estimating the treatment effect. Here, matching is an analytic as opposed to design-based strategy. 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 not implied by the experiment itself. 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 in 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. Moreover, a regression adjustment may not correct an imbalance. Matching techniques allow the analyst to ensure that imbalances have been eliminated. Matching may help to avoid charges of data snooping since adjustment is separated from the analysis of outcomes. Since the only role for covariates is their inclusion in the matching model, covariates can be separated from the actual estimation and testing of treatment effects. This does not preclude data snooping but makes it more difficult.

Matching does preserve the nonparametric nature of experiments, but it raises questions of its own. Matching can be done with and without replacement. When matching is done with replacement, each unit from the control group may be matched

to more than one treated unit. This has the potential to increase the quality of the matches and reduce bias. When matching is done without replacement each control unit is matched once to a single treated unit and used in additional matches. Bowers (2010) argues that matching with replacement hews closest to the original experimental design. We argue that either method alters the basic nature of the experiment, and it is not obvious which is to be preferred. First, consider matching without replacement. In experiments where the cells are unbalanced, matching without replacement necessitates discarding some units from the study. As such if there are more treated than control units, some treated units will be discarded or vice versa. If matching is done with replacement, however, one avoids the dropping of treated observations, but control observations may be used repeatedly thus certain control observations are dropped. The advantage, however, is that if one control observation is a good match for more than one treated unit the balance post-matching will be better and bias will be reduced to a greater extent.

The best method of variance estimation remains an open question for matching with experimental data posttreatment. In the literature on matching with observational data, variance estimation has been a particular challenge. For example, Imbens and Abadie (2006) find that the bootstrap fails at variance estimation for matching estimators. Abadie and Imbens (2006) develop special variance estimators for matching models, which only apply to matching with replacement. When matching is done without replacement, Rosenbaum (2002*b*) recommends the use of randomization inference but only when accompanied by an appropriate sensitivity analysis. How might we approach variance estimation when matching with experimental data? As we argued earlier, if there is a severe imbalance in an experiment, that experiment has essentially become an observational study. If so, variance estimation would proceed as it normally would. What if this is not the case? We conjecture that Fisher-based randomization tests are an appropriate method for statistical inference for matched experimental data when the matching is done without replacement. These tests only rely on the assumption that the probability of treatment is constant within matched units. As such, we

suspect that these tests will continue to provide correct inferences even with matching. Formal proof of our conjecture is an opportunity for future work. See Keele (2008) for a detailed treatment of Fisher-based randomization tests. When matching is done with replacement, however, variance estimation must rely on a large sample approximation, which may not be advisable for some smaller experiments. As we demonstrate in the empirical example below, the type of matching can produce very different inferences.

### 3.3 An Analytic Plan for Experimental Analysis

Before we explore two empirical examples, we ask: why adjust experimental data at all? The difference between the average response in the treatment group and the average response in the control group, often called the intention-to-treat (ITT) estimate, may be the best statistical summary of an experiment. The ITT estimate is often an acceptable estimation strategy even in the face of some forms of noncompliance. While observed imbalances may imply that the ITT estimate is biased, we would argue that the ITT estimate should always be reported. The ITT estimate need not be the average across treatment and control; it could be a comparison of medians or the Hodges-Lehmann point estimate from a rank based test, but the unadjusted estimate provides a useful summary of the experiment.

A conservative approach would suggest only reporting the ITT estimate and never adjusting the data since any adjustment method for experimental data risks unbalancing the unobservables. This critique applies to adjustments via either matching or regression. Advocates of adjustment would point out that unobservables are just that unobserved. It might be the case that observables are balanced by the randomization but unobservables are not. Or it might be the case that adjustments correct imbalance in the unobservables. As such, it is impossible to assess how any adjustment method may affect unobservables. Moreover, as we outlined earlier, there are legitimate reasons to adjust data especially if the analyst observes unbalanced covariates after randomization.

Is it possible then to stake out any course of action that might take seriously both

considerations? We doubt the advocates of models and those who oppose models will ever fully agree. As a compromise, we propose that all analysts should at least clarify whether their estimates are adjustment dependent or not. That is, any time adjustments are made the basic unadjusted estimate should also be reported along with the adjusted estimate. While further adjustments may be made, the results from these analyses should be contrasted with the ITT estimate which requires fewer assumptions and is unaffected by the methods used for adjustment. What is critical to know is whether the conclusions are heavily dependent on adjustment. As Freedman (2008*b*) notes if adjustments make a substantial difference then the results from the experiment should be treated with caution. This requires very little effort on the part of the analyst and is the most transparent method for reporting results. Moreover presenting an unadjusted and adjusted estimate increases transparency by allowing readers to understand how dependent the results are on the adjustment method.

To this end, we recommend the following steps as an analytic strategy. First, ITT estimates should be reported. Second, balance statistics from the experiment should be reported. The logic behind this is simple. Reporting balance statistics clarifies what is intended by adjustment. If randomization balances the data, adjustment is intended to increase the precision of the treatment effect. Here, one might use Rosenbaum's method of adjustment or regression if preferred. If randomization does not balance the data, adjustment is intended to correct this accidental imbalance. In this case, matching or if preferred regression models could then be used for matching. So long as the unadjusted estimate is reported along with any adjusted estimates, the reader can understand whether the inference is adjustment dependent or not. We would argue that this the critical point. While any number or form of adjustments might be made, what is important is to clarify to what extent the inference is dependent on these adjustments.

## 4 Empirical Examples

We present two empirical examples. For the first example, we conduct a comparative case study of methods for adjusting experimental data. To do this we conducted an experiment where we built several methodological manipulations into the design so that we can compare design and model-based adjustments. In the second example, we illustrate how matching might be used for posttreatment adjustments when blocking has not occurred.

### 4.1 A Comparative Method Experimental Design

We start with a description of 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 racial 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 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; perhaps creating an implicit racial cue. The story appeared to originate in the local newspaper. Subjects in the control condition were shown a story from the local newspaper about competition between the iPhone and the 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. The experimental manipulation, however, was simply a vehicle for comparing adjustment

methods. 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 posttreatment questionnaire. Subjects were paid \$5 on completion of the second session. We offered the financial incentive to decrease attrition among one arm of the study. We discuss attrition rates in a subsequent section.

It is possible that by asking subjects about race and crime prior to the experiment we activated these attitudes. One-third of the subjects were randomly assigned to complete a placebo pretest questionnaire, which 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 basic experimental manipulation. Figure 1 contains a diagram of the full design. Pretesting versus the placebo questionnaire forms the first arm of the study. Among the subjects that participated in the pretreatment questionnaire, half of them then participated in the basic experiment and the data from these subjects was analyzed as if we had not collected this pretreatment information. This forms the second arm of the study in Figure 1. These two arms of the study therefore have no design aspects that aid in adjustment. All adjustments for the data in these two arms was be done after the experiment is complete. For subjects in this arm of the study, we use our proposed analytic strategy. First, we check for balance. If we find that the study appears to be balanced by the randomization, we then report the ITT estimate along with an estimate based on Rosenbaum’s method. If the data is unbalanced, we could then use either regression or matching to adjust for the imbalance.

For the other half of the subjects, we used information from the pretreatment questionnaire to form matched pairs. We matched the subjects based on several routine background characteristics such as party identification, income, and pretreatment scales of racial attitudes and crime. This forms the final arm of the design in Figure 1. For the subjects in this arm of the study, covariates are built into the design and

therefore need not be included in the analysis phase. For these subjects, we estimate a difference in measure of location across treatment and control. We also calculate the power gain from the use of matched pairs.

This design allows us to make three specific comparisons. First, it allows us to assess whether pretreatment screening can contaminate later parts of the experiment. Second, it allows us to compare posttreatment methods for adjustment. Third, we compare the precision of matched pair results compared to unadjusted estimates.

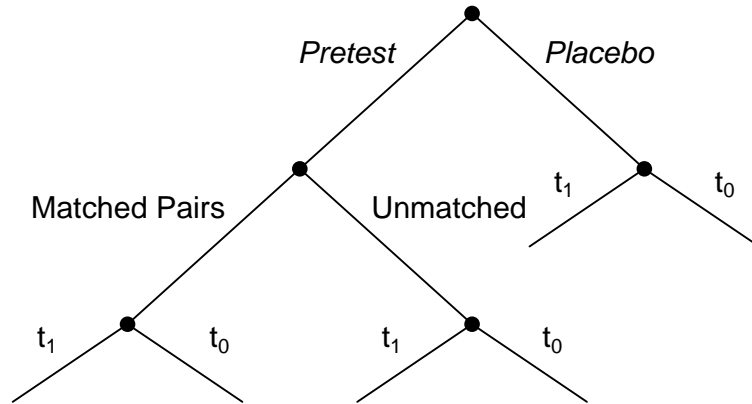


Figure 1: Comparative Method Experimental Design

## 4.2 Results

We report three different sets of results from our experiment. First, we consider whether pretreatment screening can contaminate subsequent experimental manipulations. Second, we analyze the results from matching the subjects into pairs before treatment exposure. Third, we report results from a comparison of model-based adjustment methods.

### 4.2.1 Pretreatment Contaminations

First, we investigate whether exposing subjects to questions on racial attitudes in a pretreatment questionnaire affected their responses about racial attitudes posttreatment. That is we compare responses on the racial stereotypes scale across treatment and control and across whether subjects received pretreatment racial stereotype questions. This allows us to test for two main effects and an interaction effect.<sup>4</sup> This creates two specific tests of interest. First, did the treatment effect racial attitudes? Second, is this treatment affect stronger or weaker when subjects participated in the pretreatment 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 evidence for an interaction, however, is mixed. In Figure 2, we report the standard  $2 \times 2$  plot of means for the four experimental conditions. Here, there appears to be evidence of an interaction, in that those who were exposed to pretreatment racial stereotype survey items had a stronger response to the experimental manipulation. However, we find that this interaction is not statistically significant ( $p = .406$ ). 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 pretreatment racial attitudes items. As such, we report separate results for those who received the pretreatment 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 pretreatment 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 pretreatment contamination, pretreatment screening should be avoided. A better method however is to use the design we utilized

---

<sup>4</sup>That is we can analyze the data as  $2 \times 2$  ANOVA design.

here. That is, we randomly selected some subjects to be exposed to sensitive questions in the pretreatment survey while other were not. With this design, analyst can understand the exact extent of whether any pretreatment contamination was occurred or not. Only this design based solution will fully reveal any possible contamination from pretreatment measurements.

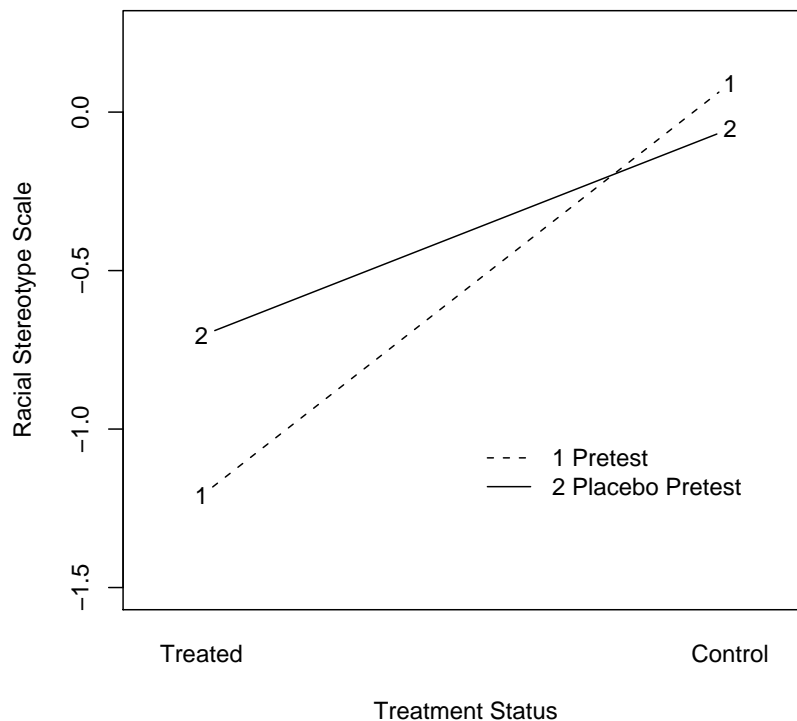


Figure 2: Interaction Between Prescreening Racial Attitudes and No Prescreening

#### 4.2.2 Pair Matching

Here, we outline how experimental analysis might proceed when covariates are built into the design. As outlined in Section 4.1, for two-thirds of the subjects, we collected a variety of covariates including the outcome measures. For half of these subjects, we used these baseline measurements to match subjects into pairs. To form the matched pairs, we used a greedy matching algorithm with a random starting point for the matches (Moore 2008). We used mahalanobis distances, and we restricted the distance

of the matches on the outcome scale to be within two points. In this process, we matched on party identification, sex, gender, race, liberal-conservative attitudes, political knowledge, as well as the outcome scale on attitudes toward crime. We purposely did not match on pretreatment racial attitudes. This allows us to observe whether failing to include covariates with match-based blocking matters to subsequent estimation of treatment effects.

Before reporting the results from this part of the study, we consider the level of attrition, since in this design, attrition must always be a concern. We lost four subjects to attrition in the matched pair arm of the study. Due the pairing of the subjects this means that four pairs or eight subjects were lost from this arm of the study reducing the effective number of pairs from 21 to 17. In this instance, then, attrition was not a severe problem though eight additional data points would be welcome given the already small sample sizes. This does highlight a concern with pre-matched designs. Care must be taken to prevent subject attrition since one missing case effectively removes other cases from the design. The advantage, however, is that once the randomization is complete and the outcomes have been collected no further adjustments are needed. Moreover, we have adjusted for these covariates in a completely nonparametric fashion. Thus this method hews very closely to the experimental ideal.

To reprise, we expect 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 1. We report the average treatment effect and use the the variance estimates derived in Imai (2008) to calculate the appropriate standard errors under the matched pair design. We find the treatment effects are in the expected direction. That is treated subjects were more anxious about crime and rated African-Americans 1.47 points lower in relation to whites on the stereotype outcomes. The  $p$ -value for the racial stereotypes outcome is somewhat higher than the usual 0.05 threshold for statistical significance, while the  $p$ -value for the crime outcome is below the 0.05 threshold ( $p = 0.02$ ). These results, however, provide little insight into whether anything was gained by matching before

the treatment.

Table 1: Matched Pairs Results

	Racial Stereotypes	Attitudes Toward Crime
ATE	-0.47	1.47
Pairs SE	0.45	0.65
Pairs $p$ -value	0.15	0.02
Unmatched SE	0.53	1.15
Unmatched $p$ -value	0.19	0.10
Efficiency Gain (%)	42	308
N	34	34

To gain insight into whether any power is gained by the pair matching, we estimated the variances for the average treatment effects ignoring the paired structure of the experiment. We report these standard errors and  $p$ -values that correspond to the unpaired variance estimates in the fourth and fifth rows of Table 1. We also report the relative efficiency—the ratio of the variance estimates—in the last row of this table. The gains in power are considerable in both cases, but much larger for the crime outcome. The use of matched pairs causes the variance estimates to be 308% smaller for the crime outcome and 43% smaller for the race outcome. The more modest efficiency gains for the race outcome may be due to the fact that we did not use measures on racial attitudes to form the matched pairs. This suggests that it is important to match on all outcomes of interest prior to treatment. Analysts must decide whether these gains in power are worth the extra difficulties created by prescreening and attrition, but this method does eliminate the need for adjustment during the analysis. Here the role of adjustment is transparent and principled as it is part of the experimental design.

### 4.2.3 Posttreatment Adjustments

As we outlined in the overview of adjustment practices, analysts rarely match before treatment. Instead adjustments are typically done with regression models after the experiment is complete. In this section, we consider how different adjustment methods might be used with the data from our experiment. We present the results in two stages.

First, we outline the basic results from the experiment. Second, we demonstrate one of the hazards of using regression models to analyze the data from experiments.

The first step in the analysis is to check balance among the covariates that we collected prior to treatment. The randomization should balance measured covariates but may not since the balancing property of randomization only holds in expectation. Any unbalanced covariates are possible candidates for statistical adjustment. We assessed balance using a randomization based test (Hansen and Bowers 2008). We present the standardized difference across treatment and control and global balance test in Table 2.<sup>5</sup> The global  $\chi^2$  test ( $p = 0.873$ ) and none of the individual standardized differences reach conventional levels of statistical significance, which indicates that the randomization successfully balanced the treatment and control groups.

Table 2: Balance Test for Crime and Race Experiment

	Standardized Difference
Party Identification	0.19
Sex	0.08
Race	-0.21
Age	-0.06
Liberal - Conservative	-0.16
Knowledge	-0.05
Income	-0.08
Frequency Watch TV News	-0.16
Political Interest	0.31
Gobal $\chi^2$ Value	6.42

Given that the data are balanced, should one then use statistical adjustments? It is possible that there may be some efficiency gains if we adjust the data. When the data are balanced as they are here, Rosenbaum’s method is perfectly suitable form of adjustment. In what follows, we report separate estimates for the subjects that were exposed to the placebo questionnaire and those who were exposed to the full pre-test questionnaire. As the reader will notice, in some instances noticeable differences arise.

<sup>5</sup>The standardized difference is simply the difference in means across treatment and control divided by a pooled standard error.

Table 3: Comparison of Estimation Methods For Racial Stereotypes

	Racial Stereotypes Placebo	Racial Stereotypes Pretest
Unadjusted Rank Sum Test	-0.50 0.109	-1.5 0.004
Rosenbaum Covariate Adjustment	-0.84 0.057	-0.35 0.21

Note: First entry is Hodges-Lehmann point estimate or mean shift.  
Second entry is exact  $p$ -value

Table 3 contains the result for the racial stereotypes outcome. The results in Table 3 are instructive in two ways. For those subjects in the placebo group, we see witness the usual logic for adjustment. By adjusting for a number of baseline covariates the variance in the outcome is reduced and we are better able to detect a treatment effect. This efficiency gain is directly observable as the exact  $p$ -value which is cut almost in half. The results for the pretest group, however, demonstrates that statistical adjustment of the data does not always lead to what we might expect. Here, the statistical adjustment which is meant to increase our power to detect a treatment effect actually reduces our precision. This can occur in highly balanced data where the regression adjustment overfits the data. One could use a penalized regression model at the adjustment stage to avoid such overfitting, but it is much simpler to report the unadjusted estimate.

Table 4: Comparison of Estimation Methods For Attitudes About Crime

	Crime Attitudes Placebo	Crime Attitudes Pretest
Unadjusted Rank Sum Test	1.5 0.048	0.75 0.268
Rosenbaum Covariate Adjustment	1.55 0.026	0.35 0.272

Note: First entry is Hodges-Lehmann point estimate or mean shift.  
Second entry is exact  $p$ -value

Table 4 contains the adjusted and unadjusted estimates for the crime outcome. For

the pretest subjects, the statistical adjustment we see that adjustment does little to change our basic inference. For the placebo subjects, however, there are minor gains in efficiency. Again presentation of both the adjusted and unadjusted estimates avoids any ambiguity about the role that adjustment plays in the estimation of the treatment effects. We also note that adjustment in our example behaves exactly as Freedman (2008*b*) predicts: at times it helps and at times it hurts. Only presentation of the unadjusted estimate helps delineate which conclusions are model dependent and which conclusions are not. Next, we conduct a more conventional analysis based on multiple regression models.

Table 5: Data Mining with Experimental Data on Racial Stereotypes

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6 Stepwise Regression
Treatment	-1.38	-1.30	-1.45	-1.62	-1.48	-1.49
	0.073	0.093	0.046	0.020	0.027	0.018
Party Id.	–	-0.49	-0.57	-0.68	-0.37	–
		-0.26	0.158	0.077	0.525	
Female	–	–	2.60	2.60	1.89	1.89
			0.000	0.000	0.009	0.005
Race	–	–	-0.20	-0.25	-0.15	–
			0.501	0.391	0.592	
Age	–	–	–	-0.06	-0.015	–
				0.574	0.883	
Frequency Watch	–	–	–	-0.95	-1.13	-1.11
TV News				0.003	0.001	0.001
Ideology	–	–	–	–	0.39	-0.68
				–	0.483	0.048
Knowledge	–	–	–	–	-1.05	-0.95
					0.002	0.001
Income	–	–	–	–	-0.13	–
					0.493	
Political Interest	–	–	–	–	-0.13	–
					0.686	

Note: Cell entries are point estimates and two-sided  $p$ -values.

One danger with adjustment of any form is that data snooping—searching for a spec-

ification that provides the “best” set of hypothesis tests—is all too easy. We demonstrate this hazard in Table 5 using the racial stereotypes measure as the outcome. We start with a simple unadjusted estimated which is reported in the first column of Table 5. The reader should note that the  $p$ -value here at 0.073 is slightly above the conventional 0.05 threshold. We then simply started adding covariates to the model in no particular order. Each additional column in Table 5 presents the results when one or more covariates have been added to the model. Finally, in the last column we present the results from using a stepwise regression algorithm to select the specification. One could argue that since the data are balanced, the additional variables being added are simply increasing the precision of the treatment effect estimate. Though as the stepwise algorithm demonstrates, there does seem to be one optimal specification in terms of minimizing the  $p$ -value on the treatment effect. This exercise of course begs the question: why should some of these covariates be included over others? Are the changes in the treatment effect  $p$ -value due to efficiency gains or incorrect variance estimates from the regression model? One could argue that the principled method would be to include all possible baseline covariates to avoid any possible data snooping. Experimental protocols, however, are typically not public and thus one often cannot know the full set of measured baseline covariates. In such instances, one can select a “best” specification such that conclusions from the data are now completely model dependent. Of course, the  $p$ -values here should also be corrected given that multiple tests have occurred, but such results from the “best” specification are typically presented without such corrections.

In this section we have focused on data from one specific experiment, however, the implications here are broadly representative. When data are balanced one might hope to gain greater precision via statistical adjustment. These adjustments, however, can move the estimated precision in either direction. Moreover, when regression models are used data snooping becomes an open possibility. The results in Table 5 demonstrated how the estimated precision for the treatment effect fluctuated with the specification. Thus adjustment should not be a automatic data analytic strategy for experimental

data. If adjustment is done, unadjusted results and balance tests should always accompany the adjusted result so that amount of model dependency is clear. We now turn to a second example to demonstrate a different possible adjustment strategy.

### 4.3 Post Hoc Experimental Analysis: Support for the Iraq War

In our second example, we analyze data from White (2003). In this example, adjustments were included in the experimental design, and we explore the options for adjustment in this context. White (2003) 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, but he manipulated the source of the article. More specifically, the same 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 their feelings on aspects of the Iraq war. We focus on two seven-point scales. The first asked whether the U.S. had tried hard enough to reach a diplomatic solution or whether he or she felt the military had become involved too quickly. The second asked whether he or she felt Iraq posed an imminent threat to the United States. 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 that 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 the results for black respondents.

The subject pool was a mix of students and adults recruited on and around Midwestern and Southern universities. Here, we might more readily expect an imbalance given that the subject pool is more heterogenous. 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 a home. We present the results from the balance tests in Table 6. Four covariates were not balanced by the randomization, and the global balance test indicates a statistically significant level of imbalance. Given this imbalance, adjustment would seem to be a logical strategy. Here, the reason for adjustment is not to increase precision but to correct for possible bias due to the existing imbalance which may be due to accidental imbalance or some error in randomization.

Of course, ANCOVA would be the usual adjustment strategy for this data. As we have reiterated, regression adjustments do not preserve the nonparametric nature of the experiment since it imposes a linear and additive functional form assumption on the data. Here, we use matching instead as the method of adjustment. As such, we report unadjusted and adjusted estimates from matching for the Iraq experiment. We used matching without and without replacement using a genetic algorithm (Sekhon and Diamond 2005; Sekhon 2007). The first step in the analysis is to test whether matching improves the balance obtained from randomization alone. Table 6 also includes the balance test results for the matched data. If we match with replacement all imbalance is removed. None of the covariates now display statistically significant imbalance and the global test is no longer statistically significant. This is not true for when we match without replacement. Here, we see that some of the imbalance is corrected but not all of it. One advantage to using matching for adjustment is any remaining imbalance should be transparent. Adjustment via regression may not correct the existing imbalance, and the remaining imbalance will not be immediately obvious to the analyst.

Now that we have corrected—or partially corrected in one case—the imbalance, we estimate treatment effects. Despite the existing imbalance in the data, we still advocate reporting unadjusted estimates along with the adjusted estimates. Anytime adjustment is necessary, the role of adjustment should be transparent and reporting the unadjusted estimates along with the adjusted estimates accomplishes this. Table 7 contains both the adjusted and unadjusted estimates for the Iraq war attitudes experiment. For each outcome, we report tests of both the sharp null hypothesis and average treatment

Table 6: Balance Tests for Iraq War Experiment

	Standardized Bias Unmatched	Standardized Bias Matched With Replacement	Standardized Bias Matched Without Replacement
Party Identification	-0.25	0.07	0.13
Liberal - Conservative	-0.38*	0.12	0.36*
Knowledge	0.37*	-0.11	-0.28
Related to Armed Services Personnel	-0.13	-0.04	0.11
Sex	-0.19	0.04	0.15
Married	-0.17	0.01	0.25
Education	0.06	-0.04	-0.05
Employed	-0.43*	0.00	0.33
Income	0.55*	-0.04	-0.53*
Own House	-0.15	-0.11	0.14
Gobal $\chi^2$ Value	22.77*	1.32	18.8*

\*  $p$ -value < 0.05

effects. To test the sharp null, we use the signed rank test for tests of the null hypothesis and the associated Hodges-Lehmann point estimate (Hollander and Wolfe 1999). The signed rank test is similar to the more familiar rank sum test but is designed for paired data. For the test of the average treatment effects estimate, we simply use a standard  $t$ -test. With the matched data we use two different methods of variance estimation. When we match without replacement, we use exact methods. When we match with replacement we use the usual Abadie-Imbens standard errors. For the first outcome in the left column, adjustment appears to matter but not a great deal. The point estimate increases from -0.99 to -1.25 and -1.50, and the  $p$ -value is smaller: 0.05 as compared to 0.038 and 0.005. Here the form of matching did alter the inference but not radically. Matching without replacement requires us to discard a small number of observations since the experiment is slightly unbalanced. There are 57 treated subjects and 60 control subjects; thus the matching algorithm discarded three control subjects. For the second outcome, however, we observe that adjustment changes our inference dramatically depending on the form of matching. The unadjusted estimate is for all

intents and purposes zero with a very small point estimate and large  $p$ -value. When we match with replacement the point estimate is large and the  $p$ -value is now below the standard 0.05 threshold ( $p = 0.039$ ). Here, of the 60 possible controls only 30 are used with one control unit being matched five different times.<sup>6</sup> When we match without replacement, the inference is much closer to the unadjusted estimate, but as is clear from Table 6 the data are imbalanced even with adjustment and thus the treatment effect estimate may be biased. This demonstrates why one should always report both the adjusted and unadjusted estimates. In this example, it allows the reader to understand that the result for the second outcome is adjustment dependent. This also illustrates that matching is no panacea. It entails its own set of choices that may alter our inference. Again transparency would appear to be the best strategy.

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

	Iraq Outcome 1	Iraq Outcome 2
Unadjusted Estimates		
Rank Sum Test	-0.99	-0.00003
	0.054	0.822
$t$ -test	-0.78	0.038
	0.051	0.907
Adjusted Estimates		
Signed Rank Test With Replacement	-1.25	-1.25
	0.031	0.039
$t$ -test With Replacement	-1.28	-0.57
	0.038	0.15
Signed Rank Test W/o Replacement	-1.50	-0.25
	0.005	0.949
$t$ -test W/o Replacement	-0.75	-0.05
	0.033	0.865
Note: First cell entry is point estimate. Second cell entry is exact or asymptotic $p$ -value.		

<sup>6</sup>We further investigated why such a dramatic change occurs. We found that the imbalance on income caused the discrepancy. A few control subjects with very high incomes reported very high scores on the outcome. This caused the difference in medians to be essentially zero.

## 5 Conclusion

Adjustment of experimental data is part of regular statistical practice in political science. The logic behind such adjustment is to either increase the precision of the estimated treatment effect or to correct for imbalances not eliminated by randomization. Currently analysts in political science rely heavily on regression model-based adjustments. In general, regression based methods of adjustment require a series of strong assumptions that are not required with experimental data. Moreover, regression models also tend to encourage poor habits with experimental data. First, it tempts the researcher to data snoop with covariates. Second, regression does not readily allow the researcher to observe whether actual imbalances have been corrected.

We outlined a number of alternatives to standard regression models. The first standard blocking or matching-based blocking builds the adjustment into the design of the experiment itself. From a statistical standpoint, this method is to be preferred. In our experiment, the pair matched design appeared powerful and did not require any adjustment. As we demonstrated, however, this was only true when we correctly pre-matched on relevant characteristics. Pair matching did not appear to improve power nearly as much for the racial outcome measure when we failed to match on pretreatment racial stereotype measures. Thus analysts should carefully consider what characteristics should be measured at baseline, so one can then match on these measures. The drawbacks to matching before the experiment are twofold. Matching requires pre-screening and may induce additional attrition. Moreover the effects of attrition are more severe since the matched units must all be present for the posttreatment analysis to occur. In some settings, it may not be possible to pre-screen. Currently, most survey experiments, for example, preclude the ability to match units prior to treatment. Second, for certain sensitive topics measurement before the treatment may itself contaminate later responses. Allowing adequate time between the measurement of baseline covariates and treatment would be important in such situations. Or analysts can use a design that allows them to test for contamination.

When adjustment seems necessary due to imbalances, standard matching techniques provide an alternative to the usual model based approaches. Matching imposes fewer assumptions on the data and preserves the nonparametric quality of the experiment. We found in the Iraq data that adjusting through matching corrected the imbalances. Matching, however, creates its own set of complications. While we conjecture that randomization inference is the correct inferential method, this has not been directly investigated. Second, in imbalanced designs some units in the study will have to be dropped. As we saw in the Iraq study this can change the results in dramatic fashion.

Randomization is perhaps the most powerful statistical tool available to applied researchers. Only with randomization, can the role of hidden confounders be minimized. Given the power of randomization, it would seem a waste to obscure the role of the treatment due to either unnecessary adjustment or an inappropriate adjustment method. While there may not be one estimation method for experimental data that will work in all instances, we argue that so long as researchers work to maximize transparency in their data analysis much will be gained. We outlined some simple suggestions that will increase transparency regardless of what method of adjustment is used. First, all experimental analyses should report balance tests. Presenting the results from balance tests allows readers to understand the role of the adjustment. When units are balanced, the case of adjustment is one of increasing precision. When units are imbalanced after the randomization, then the case for adjustment becomes more obvious. None of this is obvious if balance is not reported. Second, analysts should always report unadjusted estimates along with any adjusted estimates. Presenting both estimates maximizes transparency by showing the exact role that adjustment plays in the study conclusions. Our basic advice does not require special software or complex estimators. It simply a strategy to respect randomization.

## 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  $\alpha$  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  $\alpha$  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

- Abadie, Alberto, and Guido Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74(January): 235–267.
- Barnow, B.S., G.G. Cain, and A.S. Goldberger. 1980. "Issues in the Analysis of Selectivity Bias." In *Evaluation Studies*, ed. E. Stromsdorfer, and G. Farkas. Vol. 5 San Francisco, CA: Sage.
- Bowers, Jake. 2010. "Making Effects Manifest in Randomized Experiments." In *Cambridge Handbook of Experimental Political Science*, ed. James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. Cambridge, MA: Cambridge University Press.
- 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* 2(March): 179–196.
- Freedman, David A. 2008c. "Randomization Does Not Justify Logistic Regression."
- Greene, William H. 2000. *Econometric Analysis*. New York: Macmillan.
- 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.
- Hollander, Myles, and Douglas A. Wolfe. 1999. *Nonparametric Statistical Methods*. 2nd ed. New York, NY: John Wiley and Sons.
- Imai, Kosuke. 2008. "Variance Identification and Efficiency Analysis in Randomized Experiments Under the Matched-Pair Design." *Statistics in Medicine* 27(October): 4857–4873.
- 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.
- Imbens, Guido W., and Alberto Abadie. 2006. "On The Failure of the Bootstrap for Matching Estimators." *Econometrica* Forthcoming.
- Keele, Luke. 2008. *Semiparametric Regression for the Social Sciences*. Chichester, UK: Wiley and Sons.

- Keele, Luke, Corrine McConnaughey, and Ismail White. 2008. "Statistical Inference for Experimental Data."
- Manski, Charles F. 2007. *Identification For Prediction And Decision*. Cambridge, Mass: Harvard University Press.
- Moore, Ryan T. 2008. "Blocking to Improve Political Science Experiments."
- 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. 1989. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association* 84(December): 1024–1032.
- 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.
- Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of Propensity Scores in Observational Studies for Causal Effects." *Biometrika* 76(April): 41–55.
- 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."
- White, Ismail K. 2003. "Racial Perceptions of Support for the Iraq War" PhD thesis University of Michigan.