Matching Estimators of Causal Effects : Prospects and Pitfalls in Theory and Practice

Stephen L. Morgan and David J. Harding

Sociological Methods & Research 2006 35: 3
DOI: 10.1177/0049124106289164

Sociological Methods & Research
Volume 35 Number 1
August 2006 3-60

10.1177/0049124106289164
http://smr.sagepub.com
hosted at
http://online.sagepub.com

Matching Estimators of Causal Effects

Prospects and Pitfalls in Theory and Practice

Stephen L. Morgan
Cornell University, Ithaca, NY
David J. Harding
University of Michigan, Ann Arbor

As the counterfactual model of causality has increased in popularity, sociologists have returned to matching as a research methodology. In this article, advances over the past two decades in matching estimators are explained, and the practical limitations of matching techniques are emphasized. The authors introduce matching methods by focusing first on ideal scenarios in which stratification and weighting procedures warrant causal inference. Then, they discuss how matching is often undertaken in practice, offering an overview of the most prominent data analysis routines. With four hypothetical examples, they demonstrate how the assumptions behind matching estimators often break down in practice. Even so, the authors argue that matching techniques can be used effectively to strengthen the prosecution of causal questions in sociology.

Keywords: *matching methods; matching techniques; stratification; causal effects*

The counterfactual, or "potential outcomes," model of causality offers new possibilities for the formulation and investigation of causal questions in sociology. In the language of Holland (1986), the counterfactual perspective shifts attention from the identification of the "causes of effects" toward the more tractable goal of estimating the "effects of causes." Accordingly, the primary goal of causal analysis becomes the investigation of selected effects of a particular cause, rather than the search for all

Authors' Note: Address all correspondence to Stephen L. Morgan, Department of Sociology, 358 Uris Hall, Cornell University, Ithaca, NY 14850 (slm45@cornell.edu) or David J. Harding, Population Studies Center, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106-1248 (dharding@umich.edu).
We thank Sascha Becker, Ben Hansen, Edwin Leuven, Elizabeth Stuart, Yu Xie, the editor, and three anonymous reviewers for their detailed and extremely helpful suggestions.

possible causes of a particular outcome along with the comprehensive estimation of all of their relative effects.

The rise of the counterfactual model to prominence has increased the popularity of data analysis routines that are most clearly useful for estimating the effects of causes. The matching estimators that we review and explain in this article are perhaps the best example of a classic technique that has reemerged in the past two decades as a promising procedure for estimating causal effects. Matching represents an intuitive method for addressing causal questions, primarily because it pushes the analyst to confront the process of causal exposure as well as the limitations of available data. Accordingly, among social scientists who adopt a counterfactual perspective, matching methods are fast becoming an indispensable technique for prosecuting causal questions, even though they usually prove to be the beginning rather than the end of causal analysis on any particular topic.

Yet while empirical examples that demonstrate the potential utility of matching methods are accumulating, the methodological literature has fallen behind in providing an up-to-date treatment of the fundamentals of matching, the recent developments in practical matching methodology, and sober assessments of the strengths and weaknesses of the techniques. The purpose of this article is to provide a starting point for those sociologists who are sophisticated users of other quantitative methods and who want to understand matching methods, either to consider using matching methods in their own work or to teach matching methods in graduate methods courses. Although our agenda is primarily explanatory, we also make the case that matching techniques should be used with considerable caution and not to the exclusion of other more established methods. Even so, we see considerable promise in the usage of matching techniques to strengthen the prosecution of causal questions in sociology.

We begin with a brief discussion of the past use of matching methods. Some sociologists may be surprised to learn that the sociological literature contains some of the early developments in matching methodology. We then outline the key ideas of the counterfactual model of causality, with which most matching methods are now motivated. Then, we present the fundamental concepts underlying matching, including stratification of the data, weighting to achieve balance, and propensity scores. Thereafter, we discuss how matching is usually undertaken in practice, including an overview of various matching algorithms. Finally, we discuss how the assumptions behind matching estimators often break down in practice, and we present some of the remedies that have been proposed to address the resulting problems.

In the course of presentation, we offer four hypothetical examples that demonstrate some of the essential claims of the matching literature, progressing from idealized examples of stratification and weighting to the implementation of alternative matching algorithms on simulated data where the treatment effects of interest are known by construction. As we offer these examples, we add real-world complexity to demonstrate how such complexity can rapidly overwhelm the power of the techniques. We adopt this strategy to move beyond the sanguine methodological literature on matching, which has insufficiently demonstrated the particular weaknesses of matching techniques.

Origins and Motivations for Matching

Matching techniques have origins in experimental work from the first half of the twentieth century. Relatively sophisticated discussions of matching as a research design can be found in early methodological texts in sociology (see Greenwood 1945) and also in attempts to adjudicate between competing explanatory accounts in applied demography (Freedman and Hawley 1949). This early work continued in sociology (e.g., Althauser and Rubin 1970, 1971; Yinger, Ikeda, and Laycock 1967), right up to the key foundational literature in statistics (Rubin 1973a, 1973b, 1976a, 1976b, 1977, 1979, 1980) that provided the conceptual foundation for the new wave of matching techniques that we present in this article.

In the early 1980s, matching techniques, as we conceive of them now, were advanced in a set of papers by Rosenbaum and Rubin (1983a, 1984, 1985a, 1985b) that offered solutions to a variety of practical problems that had limited matching techniques to very simple applications in the past. Variants of these new techniques found some use immediately in sociology (Berk and Newton 1985; Berk, Newton, and Berk 1986; Hoffer, Greeley, and Coleman 1985), continuing with work by Smith (1997). In the late 1990s, economists joined in the development of matching techniques in the course of evaluating social programs (e.g., Heckman, Ichimura, and Todd 1997, 1998; Heckman, Ichimura, Smith, and Todd 1998; Heckman, LaLonde, and Smith 1999). New sociological applications are now accumulating (DiPrete and Engelhardt 2004; DiPrete and Gangl 2004; Harding 2003; Morgan 2001), and we expect that matching will complement other types of modeling in sociology with greater frequency in the future.

In the methodological literature, matching is usually introduced in one of two ways: (1) as a method to form quasi-experimental contrasts by

sampling comparable treatment and control cases from among two larger pools of such cases or (2) as a nonparametric method of adjustment for treatment assignment patterns when it is feared that ostensibly simple parametric regression estimators cannot be trusted.

For the first motivation, the archetypical example is an observational biomedical study where a researcher is called upon to assess what can be learned about a particular treatment. The investigator is given access to two sets of data, one for individuals who have been treated and one for individuals who have not. Each data set includes a measurement of current symptoms, Y_i, and a set of characteristics of individuals, as a vector of variables $\mathbf{X}_i$, that are drawn from demographic profiles and health histories. Typically, the treatment cases are not drawn from a population via any known sampling scheme. Instead, they emerge as a result of the distribution of initial symptoms, patterns of access to the health clinic, and then decisions to take the treatment. The control cases, however, may represent a subsample of health histories from some known data set. Often, the treatment is scarce, and the control data set is much larger than the treatment data set.

In this scenario, matching is a method of strategic subsampling from among treated and control cases. The investigator selects a nontreated control case for each treated case based on the characteristics in $\mathbf{X}_i$. All treated cases and matched control cases are retained, and all nonmatched control cases are discarded. Differences in Y_i are then calculated for treated and matched cases, with the average difference serving as the treatment effect estimate for the group of individuals given the treatment.[1]

The second motivation has no archetypical substantive example, as it is similar in form to any attempt to use regression to estimate causal effects with survey data. Suppose, for a general example, that an investigator is interested in the causal effect of an observed dummy variable, D_i, on an observed outcome, Y_i. But, it is known that a simple bivariate regression, $Y_i = \alpha + \gamma D_i + \varepsilon_i$, will yield an estimated coefficient $\hat{\gamma}$ that is a biased and inconsistent estimate of the causal effect of interest because the causal variable D_i is associated with variables embedded in the error term, ε_i. For a particular example, if D_i is the receipt of a college degree and Y_i is a measure of economic success, then the estimate of interest is the causal effect of obtaining a college degree on subsequent economic success. However, family background variables are present in ε_i, which are correlated with D_i, and this relationship produces classical omitted variables bias for a college degree coefficient estimated from a bivariate ordinary least squares (OLS) regression of Y_i on D_i.

In comparison to the biomedical example just presented, this motivation differs in two ways: (1) in most applications of this type, the data represent a random sample from a well-defined population, and (2) the common practice in the applied literature is to use regression to estimate effects. For the education example, a set of family background variables in a vector $\mathbf{X}_i$ is assumed to predict both D_i and Y_i. The standard regression solution is to estimate an expanded regression equation: $Y_i = \alpha + \gamma D_i + \beta' \mathbf{X}_i + \varepsilon_i$. With this strategy, the goal is to estimate simultaneously the causal effects of $\mathbf{X}_i$ and D_i on the outcome, Y_i, which may be possible because the sample is randomly drawn from a known population.

In contrast, a matching estimator nonparametrically balances the variables in $\mathbf{X}_i$ across D_i solely in the service of obtaining the best possible estimate of the causal effect of D_i on Y_i. The most popular technique is to estimate the probability of D_i for each individual i as a function of $\mathbf{X}_i$ and then to select for further analysis only matched sets of treatment and control cases that contain individuals with equivalent values for these predicted probabilities. This procedure results in a subsampling of cases, comparable to the matching procedure described for the biomedical example, but for a single dimension that is a function of the variables in $\mathbf{X}_i$. In essence, the matching procedure throws away information from the joint distribution of $\mathbf{X}_i$ and Y_i that is unrelated to variation in the treatment variable D_i until the remaining distribution of $\mathbf{X}_i$ is equivalent for both the treatment and control cases. When this equivalence is achieved, the data are said to be balanced with respect to $\mathbf{X}_i$. Under maintained assumptions that we will introduce later, the remaining differences in the observed outcome between the treatment and matched control cases can then be regarded as attributable solely to the effect of the treatment.

For the remainder of this article, we will adopt this second scenario, as research designs in which data are drawn from random-sample surveys are much more common in sociology. Thus, we will assume that the data in hand were generated by a relatively large random-sample survey, where the proportion and pattern of individuals who are exposed to the cause are fixed in the population by whatever process generates causal exposure. Moreover, we will assume for our presentation that the variables in the data are measured without error.[2]

Counterfactuals and Causal Effects

Although matching can be seen as an extension of the tabular analysis of simple three-way cross-classifications (i.e., outcome variable by causal

variable by adjustment variable), the current literature is primarily associated with counterfactual models of causality.[3] Accordingly, from here onward, we adopt the language that dominates this framework, writing of the causal variable of interest as a treatment variable. And, as will become apparent latter, we confine most of our attention to binary treatments, generally referring to the group that receives the treatment as the treatment group and the group that does not as the control group.[4] One could, however, rewrite all that follows referring to such groups as those who are exposed to the cause and those who are not.

Causal Effects of Primary Interest

In the counterfactual framework, we approach causal inference by first stipulating the existence of two potential outcome random variables that are defined over all individuals in the population. Y_i^1 is the potential outcome in the treatment state for individual i, and Y_i^0 is the potential outcome in the control state for individual i. The individual-level causal effect of the treatment is then defined as

$$\delta_i = Y_i^1 - Y_i^0. \tag{1}$$

Because we can never observe the potential outcome under the treatment state for those observed in the control state (and vice versa), we can never know the individual-level causal effects in equation (1).[5] This predicament is sometimes labeled the fundamental problem of causal inference (Holland 1986). Instead, we can only observe values for a variable Y_i, which is related to the potential outcomes of each individual by

$$Y_i = Y_i^1 \text{ if } D_i = 1,$$
$$Y_i = Y_i^0 \text{ if } D_i = 0,$$

where the binary variable, D_i, is equal to 1 if an individual receives the treatment (i.e., is exposed to the cause) and equal to 0 if an individual receives the control (i.e., is not exposed to the cause). This paired definition is generally written compactly as

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0. \tag{2}$$

Because it is usually impossible to effectively estimate individual-level causal effects, we typically shift attention to aggregated causal effects. With $E[.]$ denoting the expectation operator from probability theory, the average causal effect is

$$E[\delta_i] = E[Y_i^1] - E[Y_i^0]. \tag{3}$$

For equation (3), the expectation is defined with reference to the population of interest, and the conditioning on i is redundant (because the causal effect of a randomly selected individual from the population is equal to the average causal effect across individuals in the population). Nonetheless, as with most other work in the counterfactual framework, we will preserve conditioning on i in our notation for causal effects, as it reinforces the inherent individual-level heterogeneity of potential outcomes and causal effects.

Although the unconditional average treatment effect is the most common subject of investigation in sociology, more narrowly defined average treatments are of interest as well, as we show in the examples later. The average treatment effect for those who take the treatment is

$$E[\delta_i|D_i=1]=E[Y_i^1|D_i=1]-E[Y_i^0|D_i=1], \tag{4}$$

and the average treatment effect for those who do not take the treatment is

$$E[\delta_i|D_i=0]=E[Y_i^1|D_i=0]-E[Y_i^0|D_i=0]. \tag{5}$$

As will become clear, in many cases, only one of the two average treatments effects in equations (4) and (5) can be estimated consistently, and when this is the case, the overall average treatment effect in equation (3) cannot be estimated consistently.

Other average causal effects (or more general properties of the distribution of causal effects) are often of interest as well, and Heckman (2000), Manski (1995), Rosenbaum (2002), and Pearl (2000) all give full discussions of the variety of causal effects that may be relevant for different types of applications. In this article, we focus almost exclusively on the three types of average causal effects represented by equations (3), (4), and (5).

Naive Estimation of Causal Effects

Having introduced the notation of the counterfactual model, in this section, we explain why matching may be necessary by demonstrating the general weakness of what has become known as the "naive estimator" in the literature. We also use this section to introduce notation for sample-based quantities, which can then be related to the population-level expectations of the last section, as well as the general large-sample inference framework that we use throughout our presentation.

For this section, we assume that randomization of the treatment is infeasible in the unspecified application that is under consideration. Instead, an autonomous fixed treatment selection regime prevails, where π is the proportion of the population that takes the treatment instead of the control. Thus, the value of π is fixed in the population by the behavior of individuals, and it is unknown to the researcher.

We assume that the researcher has observed survey data from a relatively large random sample of the population. For the sample expectation of a quantity in a sample of size N, we will use a subscript on the expectation operator, as in $E_N[.]$. Accordingly, $E_N[D_i]$ is the sample mean of the dummy treatment variable, $E_N[Y_i|D_i = 1]$ is the sample mean of the outcome for those observed in the treatment group, and $E_N[Y_i|D_i = 0]$ is the sample mean of the outcome for those observed in the control group.[6] The naive estimator of the average causal effect is then defined as

$$\hat{\delta}_{NAIVE} \equiv E_N[Y_i|D_i = 1] - E_N[Y_i|D_i = 0], \tag{6}$$

which is simply the difference in the sample means of the observed outcome variable Y_i for the observed treatment and control cases in the full sample (i.e., prior to any subsampling via a matching routine).

In the absence of randomization of the treatment, the naive estimator rarely yields a consistent estimate of the average treatment effect because it converges to a contrast, $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$, that is not equivalent to any of the causal effects defined in the last section. To make this clear, we can decompose the average causal effect as

$$\begin{aligned} E[\delta_i] = &\{\pi E[Y_i^1|D_i = 1] + (1 - \pi)E[Y_i^1|D_i = 0]\} \\ &- \{\pi E[Y_i^0|D_i = 1] + (1 - \pi)E[Y_i^0|D_i = 0]\}. \end{aligned} \tag{7}$$

The average treatment effect is then a function of five unknowns: the proportion of the population that is assigned to (or self-selects into) the treatment, along with four conditional expectations of the potential outcomes. Without introducing additional assumptions, we can consistently estimate with observational data from a sample of the population only three of the five unknowns on the right-hand side of equation (7).

We know that, for a very large random sample, the mean of the dummy treatment variable D_i would be equal to the true proportion of the population that would be assigned to (or would select into) the treatment. More precisely, we know that the sample mean of D_i converges in probability to π, which we write as

$$E_N[D_i] \xrightarrow{p} \pi. \tag{8}$$

Although the notation of equation (8) may appear unfamiliar, the claim is that, as the sample size N increases, the sample mean of D_i approaches the true value of π, which we assume is a fixed population parameter. Thus, the notation $\xrightarrow{p}$ connotes convergence in probability for a sequence of estimates over a set of samples where the sample size N is increasing. We can offer similar claims about two other unknowns in equation (7):

$$E_N[Y_i|D_i=1] \xrightarrow{p} E[Y_i^1|D_i=1], \tag{9}$$

$$E_N[Y_i|D_i=0] \xrightarrow{p} E[Y_i^0|D_i=0], \tag{10}$$

which indicate that the sample mean of the observed outcome in the treatment group converges to the true average outcome under the treatment state for those in the treatment group (and analogously for the control group and control state).[7]

Unfortunately, however, there is no assumption-free way to effectively estimate the two remaining unknowns in equation (7): $E[Y_i^1|D_i=0]$ and $E[Y_i^0|D_i=1]$. These are counterfactual conditional expectations: the average outcome under the treatment for those in the control and the average outcome under the control for those in the treatment. Without further assumptions, no estimated quantity based on observed data from a random sample of the population would converge to the true values for these unknowns.

Estimating Causal Effects Under Maintained Assumptions About Potential Outcomes

What assumptions would suffice to enable consistent estimation of the average treatment effect with observed data? There are two basic classes of assumptions, which can be regarded as mirror images of each other: (1) assumptions about potential outcomes for subsets of the population defined by treatment status and (2) assumptions about the treatment assignment/selection process in relation to the potential outcomes. For the first type of assumption, we could assert the following two equalities:

$$\text{Assumption 1:} \quad E[Y_i^1|D_i=1] = E[Y_i^1|D_i=0], \tag{11}$$

$$\text{Assumption 2:} \quad E[Y_i^0|D_i=1] = E[Y_i^0|D_i=0], \tag{12}$$

and then substitute into equation (7) to reduce the number of unknowns from the original five parameters to the three parameters that we know from equations (8) through (10) can be consistently estimated with the data.

For the second type of assumption, we could assert what has become known as an assumption of the ignorability of treatment assignment (see Rubin 1978). If the treatment D_i is completely independent of the potential outcomes Y_i^0 and Y_i^1 (as well as any function of them, such as the distribution of δ_i), then treatment assignment is ignorable.[8] When treatment assignment is ignorable in this sense, Assumptions 1 and 2 are implied.

For our presentation, asserting assumptions about average differences in potential outcomes for subsets of the population, as in Assumptions 1 and 2, is often more straightforward than invoking assumptions about the independence of treatment assignment from potential outcomes. Consider the following two scenarios that demonstrate why, where first we note the utility of ignorability assumptions in randomization research designs.

If D_i is assigned by some completely random process in the population, then treatment assignment is ignorable because D_i is fully independent of everything defined on the population before the treatment is applied (i.e., Y_i^0, Y_i^1, δ_i, etc.). Random treatment assignment justifies Assumptions 1 and 2 since, by the same reasoning, the average difference between those in the treatment group and those in the control must be zero for both Y_i^0 and Y_i^1 if full independence of D_i is assumed. But it is somewhat more natural to discuss the implications of randomization via ignorability assumptions, as the randomization operation is the most prominent feature of the research design. Accordingly, ignorability assumptions are common in discussions of matching in biostatistics where randomization is widespread.

Now consider the sort of research designs that are most prevalent in the social sciences. Here, randomization is infeasible, the outcome Y_i has an inherent metric of theoretical interest, and treatment assignment is more often a process of self-selection (or nonrandom allocation) than assignment by an autonomous randomizer. For this type of research, expectation-based assumptions such as Assumptions 1 and 2 are commonly used. It is often more natural to directly assert Assumption 1 or Assumption 2 (or perhaps both) based on theoretical conjectures about the "what if" average levels of counterfactual potential outcomes for individuals if they had been exposed to an alternative cause. Backing all the way up to an all-encompassing assumption of ignorability of treatment assignment proves unnatural (and is often unnecessary if one is only interested in estimating either the average treatment effect for the treated or the average treatment effect for the untreated).

In the remaining sections of this article, we present matching estimators within the counterfactual framework we have introduced in this section.

As we noted at the outset, matching is a set of techniques that can be motivated in other ways. And, in fact, the classical matching literature is an outgrowth of experimental methodology rather than the early potential outcomes framework that now undergirds the counterfactual perspective. Nonetheless, most of the work over the past three decades on matching has been motivated with this framework for observational data analysis, and thus we follow in this tradition.

Matching as Stratification

Having covered the preliminaries, in this section, we introduce matching estimators in idealized research conditions. Thereafter, we proceed to a discussion of matching in more realistic scenarios, which is where we explain the developments of matching techniques that have been achieved in the past three decades as well as the substantial problems that remain and limit the ultimate usefulness of matching.

Estimating Causal Effects by Stratification

Suppose that those who take the treatment and those who do not are very much unlike each other, and yet the ways in which they differ are captured exhaustively by a set of observed treatment assignment/selection variables $\mathbf{S}$.[9] For the language we will adopt in this article, knowledge and observation of $\mathbf{S}$ allow for a "perfect stratification" of the data. By "perfect," we mean precisely that individuals within groups defined by values on the variables in $\mathbf{S}$ are entirely indistinguishable from each other in all ways except for (1) observed treatment status and (2) completely random shocks to the potential outcome variables. Under such a perfect stratification of the data, even though we would not be able to assert Assumptions 1 and 2, we would be able to assert conditional variants of those assumptions:

$$\text{Assumption 1-S:} \quad E[Y_i^1 | D_i = 1, \mathbf{S}_i] = E[Y_i^1 | D_i = 0, \mathbf{S}_i], \tag{13}$$

$$\text{Assumption 2-S:} \quad E[Y_i^0 | D_i = 1, \mathbf{S}_i] = E[Y_i^0 | D_i = 0, \mathbf{S}_i]. \tag{14}$$

These assumptions would suffice to enable consistent estimation of the average treatment effect, as the treatment can be considered randomly assigned within groups defined by values on the variables in $\mathbf{S}$.[10]

Before we introduce an idealized example of stratification, first note why everything works out so cleanly when a set of perfect stratifying variables

is available. If Assumption 1-S is valid, then

$$\begin{aligned} E[\delta_i|D_i=0,\mathbf{S}_i] &= E[Y_i^1|D_i=0,\mathbf{S}_i] - E[Y_i^0|D_i=0,\mathbf{S}_i] \\ &= E[Y_i^1|D_i=1,\mathbf{S}_i] - E[Y_i^0|D_i=0,\mathbf{S}_i]. \end{aligned} \tag{15}$$

If Assumption 2-S is valid, then

$$\begin{aligned} E[\delta_i|D_i=1,\mathbf{S}_i] &= E[Y_i^1|D_i=1,\mathbf{S}_i] - E[Y_i^0|D_i=1,\mathbf{S}_i] \\ &= E[Y_i^1|D_i=1,\mathbf{S}_i] - E[Y_i^0|D_i=0,\mathbf{S}_i]. \end{aligned} \tag{16}$$

Both of the last two lines of equations (15) and (16) are identical, and neither includes counterfactual conditional expectations. One can consistently estimate the differences in the last two lines of equations (15) and (16) if these assumptions hold and thus obtain consistent estimates of treatment effects conditional on **S**. To then form consistent estimates of alternative average treatment effects, one simply averages the stratified estimates over the distribution of **S**, as we show in the following hypothetical example.

Hypothetical Example 1

Consider a completely hypothetical example where Assumptions 1 and 2 cannot be asserted because positive self-selection ensures that those who are observed in the treatment group are more likely to benefit from the treatment than those who are not. But assume that a three-category perfect stratifying variable S is available that allows one to assert Assumptions 1-S and 2-S. Moreover, suppose for simplicity of exposition that our sample is large enough such that sampling error is trivial. Therefore, we can assume that the sample moments in our data equal the population moments (i.e., $E_N[Y_i|D_i=1] = E[Y_i|D_i=1]$ and so on).[11]

If it is helpful, for this example, the reader can think of Y_i as a measure of an individual's economic success at age 40, D_i as an indicator of receipt of a college degree, and S_i as a mixed family background and preparedness-for-college variable that completely accounts for the pattern of self-selection into college that is relevant for lifetime economic success. Note, however, that no one has ever discovered such a variable as S for this particular causal effect.[12] For economy of space, however, we will refer to these variables generically as S, Y, and D below.

Suppose now that, for our very large sample, the sample mean of the outcome for those observed in the treatment group is 10.2, whereas the sample mean of the outcome for those observed in the control group is

Table 1
The Joint Probability Distribution and Conditional Population Expectations for Hypothetical Example 1

	Joint Probability Distribution of S and D		
	$D=0$	$D=1$	
$S=1$	$\Pr[S=1, D=0]=.36$	$\Pr[S=1, D=1]=.08$	$\Pr[S=1]=.44$
$S=2$	$\Pr[S=2, D=0]=.12$	$\Pr[S=2, D=1]=.12$	$\Pr[S=2]=.24$
$S=3$	$\Pr[S=3, D=0]=.12$	$\Pr[S=3, D=1]=.2$	$\Pr[S=3]=.32$
	$\Pr[D=0]=.6$	$\Pr[D=1]=.4$	
	Potential Outcomes		
	Under the Control State	Under the Treatment State	
$S=1$	$E[Y^0\|S=1]=2$	$E[Y^1\|S=1]=4$	$E[Y^1-Y^0\|S=1]=2$
$S=2$	$E[Y^0\|S=2]=6$	$E[Y^1\|S=2]=8$	$E[Y^1-Y^0\|S=2]=2$
$S=3$	$E[Y^0\|S=3]=10$	$E[Y^1\|S=3]=14$	$E[Y^1-Y^0\|S=3]=4$
	$E[Y^0\|D=0]$ $=\frac{.36}{.6}(2)+\frac{.12}{.6}(6)+\frac{.12}{.6}(10)$ $=4.4$	$E[Y^1\|D=1]$ $=\frac{.08}{.4}(4)+\frac{.12}{.4}(8)+\frac{.2}{.4}(14)$ $=10.2$	

4.4. In other words, we have data that yield $E_N[Y_i|D_i=1]=10.2$ and $E_N[Y_i|D_i=0]=4.4$ and where the naive estimator would yield a value of 5.8 (i.e., $10.2-4.4$).

Consider, now, an underlying set of potential outcome variables and treatment assignment patterns that could give rise to a naive estimate of 5.8. Table 1 presents the joint probability distribution of the treatment variable D and the stratifying variable S in its first panel as well as expectations, conditional on S, of the potential outcomes under the treatment and control states. The joint distribution in the first panel shows that individuals with S equal to 1 are more likely to be observed in the control group, individuals with S equal to 2 are equally likely to be observed in the control group and the treatment group, and individuals with S equal to 3 are more likely to be observed in the treatment group.

As shown in the second panel of Table 1, the average potential outcomes conditional on S and D imply that the average causal effect is 2 for those with S equal to 1 or S equal to 2 but 4 for those with S equal to 3 (see the last column). Moreover, as shown in the last row of the table, where the potential outcomes are averaged over the within-D distribution

Table 2
Estimated Conditional Expectations and Probabilities From a Very Large Sample for Hypothetical Example 1

	Estimated Mean of the Observed Outcome Conditional on S and D	
	Control Group	Treatment Group
$S_i = 1$	$E_N[Y_i \mid S_i = 1, D_i = 0] = 2$	$E_N[Y_i \mid S_i = 1, D_i = 1] = 4$
$S_i = 2$	$E_N[Y_i \mid S_i = 2, D_i = 0] = 6$	$E_N[Y_i \mid S_i = 2, D_i = 1] = 8$
$S_i = 3$	$E_N[Y_i \mid S_i = 3, D_i = 0] = 10$	$E_N[Y_i \mid S_i = 3, D_i = 1] = 14$
	Estimated Probability of S Conditional on D	
$S_i = 1$	$\Pr_N[S_i = 1 \mid D_i = 0] = .6$	$\Pr_N[S_i = 1 \mid D_i = 1] = .2$
$S_i = 2$	$\Pr_N[S_i = 2 \mid D_i = 0] = .2$	$\Pr_N[S_i = 2 \mid D_i = 1] = .3$
$S_i = 3$	$\Pr_N[S_i = 3 \mid D_i = 0] = .2$	$\Pr_N[S_i = 3 \mid D_i = 1] = .5$

of S, $E[Y \mid D = 0] = 4.4$ and $E[Y \mid D = 1] = 10.2$, matching the initial setup of the example based on a naive estimate of 5.8 from a very large sample.

Table 2 shows what can be calculated from the data, assuming that S offers a perfect stratification of the data. The first panel presents the sample expectations of the observed outcome variable conditional on D and S. The second panel of Table 2 presents corresponding sample estimates of the conditional probabilities of S given D. The estimated values are for a very large sample, as stipulated earlier, such that sampling error is infinitesimal.

The existence of a perfect stratification (and the availability of a very large data set) ensures that the estimated conditional expectations in the first panel of Table 2 match the population-level conditional expectations of the second panel of Table 1. When stratifying by S, the average observed outcome for those in the control/treatment group with a particular value of S is equal to the average potential outcome under the control/treatment state for those with a particular value of S. Conversely, if S were not a perfect stratifying variable, then the sample means in the first panel of Table 2 would not equal the expectations of the potential outcomes in the second panel of Table 1. The sample means would be based on heterogeneous groups of individuals who differ systematically within the strata defined by S in ways that are correlated with individual-level treatment effects.

If S offers a perfect stratification of the data, then, with a suitably large sample, one can estimate from the numbers in the cells of the two panels of Table 2 both the average treatment effect among the treated as $(4-2)(.2)+(8-6)(.3)+(14-10)(.5)=3$ and the average treatment effect among the untreated as $(4-2)(.6)+(8-6)(.2)+(14-10)(.2)=2.4$. Finally, if one calculates the appropriate marginal distributions of S and D (using sample analogs for the marginal distribution from the first panel of Table 1), one can perfectly estimate the unconditional average treatment effect either as $(4-2)(.44)+(8-6)(.24)+(14-10)(.32)=2.64$ or as $3(.6)+2.4(.4)=2.64$. Thus, for this hypothetical example, the naive estimator would be asymptotically upwardly biased for the average treatment effect among the treated, the average treatment effect among the untreated, and the unconditional average treatment effect. But, by appropriately weighting stratified estimates of the treatment effect, unbiased and consistent estimates of the average treatment effects in equations (3), (4), and (5) can be obtained.

In general, if a stratifying variable S completely accounts for all systematic differences between those who take the treatment and those who do not, then conditional-on-S estimators yield consistent estimates of the average treatment effect conditional on S:

$$\{E_N[Y_i|D_i=1,S_i=s]-E_N[Y_i|D_i=0,S_i=s]\} \xrightarrow{p} E[Y_i^1-Y_i^0|S_i=s]=E[\delta_i|S_i=s].$$

One can then take weighted sums of these stratified estimators, such as for the unconditional average treatment effect:

$$\sum_S \{E_N[Y_i|D_i=1,S_i=s]-E_N[Y_i|D_i=0,S_i=s]\}\Pr{}_N[S_i=s] \xrightarrow{p} E[\delta_i].$$

Substituting into this last expression the distributions of S conditional on the two possible values of D, one can obtain consistent estimates of the average treatment effect among the treated and the average treatment effect among the untreated.

The key to using stratification to solve the causal inference problem for all three causal effects of primary interest is twofold: finding the stratifying variable and then obtaining the marginal probability distribution $\Pr(S)$ as well as the conditional probability distribution $\Pr(S|D)$. Once these steps are accomplished, obtaining consistent estimates of the within-strata treatment effects is straightforward, and one simply forms the appropriate weighted average of the stratified estimates.

This simple example shows all of the basic principles of matching estimators. Treatment and control subjects are matched together in the sense that they are grouped together into strata. Then, an average difference between the outcomes of treatment and control subjects is estimated, based on a weighting of the strata (and thus the individuals within them) by a common distribution—that is, the marginal distribution $\Pr(S)$, the conditional distribution $\Pr(S|D=1)$, the opposite conditional distribution $\Pr(S|D=0)$, or any other theoretically meaningful distribution of S. The imposition of the same set of stratum-level weights for those in both the treatment and control groups ensures that the data are balanced with respect to the distribution of S across treatment and control cases.

Overlap Conditions for Stratifying Variables

Suppose now that a perfect stratification of the data is available but that there is a stratum in which no member of the population ever receives the treatment. Here, the average treatment effect is undefined. A hidden stipulation is built into Assumptions 1-S and 2-S if one wishes to be able to estimate the average treatment effect for the entire population. The ''perfect'' stratifying variables must not be so perfect that they sort deterministically individuals into either the treatment and control. If so, the range of the stratifying variables will differ fundamentally for treatment and control cases, necessitating a redefinition of the causal effect of interest.

Hypothetical Example 2

For the example depicted in Tables 3 and 4, S again offers a perfect stratification of the data. The setup of these two tables is exactly equivalent to the prior Tables 1 and 2. The major difference is evident in the joint distribution of S and D presented in the first panel of Table 3. As shown in the first cell of the second column, no individual with S equal to 1 would ever be observed in the treatment group of a data set of any size, as the joint probability of S equal to 1 and D equal to 1 is zero. Corresponding to this structural zero in the joint distribution of S and D, the second panel of Table 3 shows that there is no corresponding conditional expectation of the potential outcome under the treatment state for those with S equal to 1. And thus, as shown in the last column of the second panel of Table 3, no causal effect exists for individuals with S equal to 1.[13]

Adopting the college degree causal effect framing of the last hypothetical example, this hypothetical example asserts that there is a subpopulation

Table 3
The Joint Probability Distribution and Conditional Population Expectations for Hypothetical Example 2

	Joint Probability Distribution of S and D		
	$D=0$	$D=1$	
$S=1$	$\Pr[S=1, D=0]=.4$	$\Pr[S=1, D=1]=0$	$\Pr[S=1]=.4$
$S=2$	$\Pr[S=2, D=0]=.1$	$\Pr[S=2, D=1]=.13$	$\Pr[S=2]=.23$
$S=3$	$\Pr[S=3, D=0]=.1$	$\Pr[S=3, D=1]=.27$	$\Pr[S=3]=.37$
	$\Pr[D=0]=.6$	$\Pr[D=1]=.4$	
	Potential Outcomes		
	Under the Control State	Under the Treatment State	
$S=1$	$E[Y^0\|S=1]=2$		
$S=2$	$E[Y^0\|S=2]=6$	$E[Y^1\|S=2]=8$	$E[Y^1-Y^0\|S=2]=2$
$S=3$	$E[Y^0\|S=3]=10$	$E[Y^1\|S=3]=14$	$E[Y^1-Y^0\|S=2]=4$
	$E[Y^0\|D=0]$ $=\frac{.4}{.6}(2)+\frac{.1}{.6}(6)+\frac{.1}{.6}(10)$ $=4$	$E[Y^1\|D=1]$ $=\frac{.13}{.4}(8)+\frac{.27}{.4}(14)$ $=12.05$	

of individuals from such disadvantaged backgrounds that no individuals with $S=1$ have ever graduated from college. For this group of individuals, we assume in this example that there is simply no justification for using the wages of those from more advantaged social backgrounds to extrapolate to the "what-if" wages of the most disadvantaged individuals if they had somehow overcome the obstacles that prevent them from obtaining college degrees.

Table 4 shows what can be estimated from a very large sample for this example. If S offers a perfect stratification of the data, one could consistently estimate the treatment effect for the treated as $(8-6)(.325)+(14-10)(.675)=3.35$. There is, unfortunately, no way to consistently estimate the treatment effect for the untreated and hence no way to consistently estimate the unconditional average treatment effect.

Are examples such as this one ever found in practice? For an example that is more realistic than the causal effect of a college degree on economic success, consider the evaluation of a generic program in which there is an eligibility rule. One simply cannot estimate the likely benefits of enrolling in the program for those who are ineligible, even though, if some of those individuals were enrolled in the program, they would likely

Table 4
Estimated Conditional Expectations and Probabilities From a Very Large Sample for Hypothetical Example 2

	Estimated Mean of the Observed Outcome Conditional on S and D	
	Control Group	Treatment Group
$S_i = 1$	$E_N[Y_i \mid S_i = 1, D_i = 0] = 2$	
$S_i = 2$	$E_N[Y_i \mid S_i = 2, D_i = 0] = 6$	$E_N[Y_i \mid S_i = 2, D_i = 1] = 8$
$S_i = 3$	$E_N[Y_i \mid S_i = 3, D_i = 0] = 10$	$E_N[Y_i \mid S_i = 3, D_i = 1] = 14$
	Estimated Probability of S Conditional on D	
$S_i = 1$	$\Pr_N[S_i = 1 \mid D_i = 0] = .667$	$\Pr_N[S_i = 1 \mid D_i = 1] = 0$
$S_i = 2$	$\Pr_N[S_i = 2 \mid D_i = 0] = .167$	$\Pr_N[S_i = 2 \mid D_i = 1] = .325$
$S_i = 3$	$\Pr_N[S_i = 3 \mid D_i = 0] = .167$	$\Pr_N[S_i = 3 \mid D_i = 1] = .675$

be affected by the treatment in some way (but, of course, in a way that may be very different from those who do enroll in the program).

Perhaps the most important point of this last example, however, is that the naive estimator is entirely misguided for this hypothetical application. The average treatment effect is undefined for the population of interest. More generally, not all causal questions have answers worth seeking even in best-case data availability scenarios, and sometimes this will be clear from the data and contextual knowledge of the application. However, at other times, the data may appear to suggest that no causal inference is possible for some group of individuals even though the problem is simply a small sample size. There is a clever solution to sparseness of data for these types of situations, which we discuss in the next section.

Matching as Weighting

As shown in the last section, if all of the variables in **S** have been observed such that a perfect stratification of the data would be possible with a suitably large random sample from the population, then a consistent estimator is available in theory for each of the average causal effects of interest defined in equations (3) through (5). However, in many (if not most) data sets of finite size, it may not be possible to use the simple estimation methods of the last section to generate consistent estimates. Treatment and control cases

may be missing at random within some of the strata defined by **S**, such that some strata contain only treatment or only control cases. In this scenario, some within-stratum causal effect estimates cannot be calculated. In this section, we introduce a related set of weighting estimators that rely on estimated propensity scores to solve the sort of data sparseness problems that afflict samples of finite size.

The Utility of Known Propensity Scores

An estimated propensity score is the estimated probability of taking the treatment as a function of variables that predict treatment assignment. Before explaining the attraction of estimated propensity scores, there is value in understanding why known propensity scores would be useful in an idealized context such as a perfect stratification of the data.

Within a perfect stratification, the true propensity score is nothing other than the within-stratum probability of receiving the treatment, or $\Pr[D=1|S]$. For hypothetical Example 1, the propensity scores are as follows:

$$\Pr[D=1|S=1]=\frac{.08}{.44}=.182,$$
$$\Pr[D=1|S=2]=\frac{.12}{.24}=.5, \text{ and}$$
$$\Pr[D=1|S=3]=\frac{.2}{.32}=.625.$$

Why is the propensity score useful? As shown earlier for hypothetical Example 1, if a perfect stratification of the data is available, then the final ingredient for calculating average treatment effect estimates for the treated and for the untreated is the conditional distribution $\Pr[S|D]$. One can recover $\Pr[S|D]$ from the propensity scores by applying Bayes's rule using the marginal distributions of D and S. For example, for the first stratum in Example 1,

$$\Pr[S=1|D=1]=\frac{\Pr[D=1|S=1]\Pr[S=1]}{\Pr[D=1]}=\frac{(.182)(.44)}{(.4)}=.2.$$

Thus, the true propensity scores encode all of the necessary information about the joint dependence of S and D that is needed to estimate and then combine conditional-on-S treatment effect estimates into estimates of the treatment effect for the treated and the treatment effect for the untreated. Known propensity scores are thus useful for unpacking the

inherent heterogeneity of causal effects and then averaging over such heterogeneity to calculate average treatment effects.

Of course, known propensity scores are almost never available to researchers working with observational rather than experimental data. Thus, the literature on matching more often recognizes the utility of propensity scores for addressing an entirely different concern: solving comparison problems created by the sparseness of data in any finite sample. These methods rely on estimated propensity scores, as we discuss next.

Weighting With Propensity Scores to Address Sparseness

Suppose again that a perfect stratification of the data exists and is known. In particular, Assumptions 1-S and 2-S are valid for a set of variables in **S**, which are measured without error. Further suppose that the true propensity score is greater than 0 and less than 1 for every possible combination of values on the variables in **S**. But suppose now that (1) there are multiple variables in **S**, and (2) some of these variables take on many values. In this scenario, there may be many strata in the available data in which no treatment or control cases are observed, even though the true propensity score is between 0 and 1 for every stratum in the population.

Can average treatment effects be consistently estimated in this scenario? Rosenbaum and Rubin (1983a) answer this question affirmatively. The essential points of their argument are the following (and see the original article for a formal proof). First, the sparseness that results from the finiteness of a sample is random, conditional on the joint distribution of **S** and D. As a result, within each stratum for a perfect stratification of the data, the probability of having a zero cell in the treatment or the control state is solely a function of the propensity score. Because such sparseness is conditionally random, strata with identical propensity scores (i.e., different combinations of values for the variables in **S** but the same within-stratum probability of treatment) can be combined into a more coarse stratification. Over repeated samples from the same population, zero cells would emerge with equal frequency across all strata within these coarse propensity score–defined strata.

Because sparseness emerges in this predictable fashion, stratifying on the propensity score itself (rather than more finely on all values of the variables in **S**) solves the sparseness problem because the propensity score can be treated as a single perfectly stratifying variable. In fact, as we show in the next hypothetical example, one can obtain consistent estimates of treatment effects by weighting the individual-level data by an appropriately

chosen function of the propensity score, without ever having to compute any stratum-specific causal effect estimates.

But how does one obtain the propensity scores for data from a random sample of the population of interest? Rosenbaum and Rubin (1983a) argue that if one has observed the variables in $\mathbf{S}$, then the propensity score can be estimated using standard methods, such as logit modeling. That is, one can estimate the propensity score, assuming a logistic distribution:

$$\Pr[D_i = 1|\mathbf{S}_i] = \frac{\exp(\mathbf{S}_i\boldsymbol{\phi})}{1 + \exp(\mathbf{S}_i\boldsymbol{\phi})} \tag{17}$$

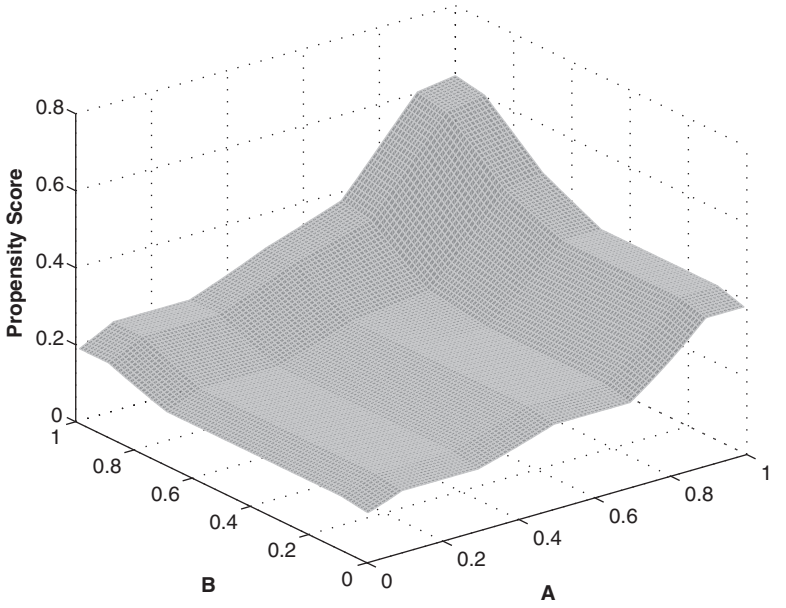
and invoke maximum likelihood to estimate the vector of coefficients $\boldsymbol{\phi}$. One can then stratify on the index of the estimated propensity score, $e(\mathbf{S}_i) = \mathbf{S}_i\hat{\boldsymbol{\phi}}$, or appropriately weight the data, as we show in the next example, and all of the results established for known propensity scores then obtain.[14] Consider the following hypothetical example, where weighting is performed only with respect to the estimated propensity score, resulting in unbiased and consistent estimates of average treatment effects even though sparseness problems are severe.

Hypothetical Example 3

Consider the following Monte Carlo simulation, which is an expanded version of hypothetical Example 1 in two respects. First, for this example, there are two stratifying variables, A and B, each of which has 100 separate values. As for Example 1, these two variables represent a perfect stratification of the data and, as such, represent all of the variables in the vector of perfect stratifying variables, defined earlier as $\mathbf{S}$. Second, to demonstrate the properties of alternative estimators, this example uses $50,000$ samples of data, each of which is a random realization of the same set of definitions for the constructed variables and the stipulated joint distribution between them.

Generation of the 50,000 data sets. For the simulation, we gave the variables A and B values of .01, .02, .03, and upward to 1. We then cross-classified the two variables to form a 100-by-100 grid and stipulated a propensity score, as displayed in Figure 1, that is a positive, nonlinear function of both A and B.[15] We then populated the resulting 20,000 constructed cells (100 by 100 for the A-by-B grid multiplied by the two values of D) using a Poisson random-number generator with the relevant propensity score as the Poisson parameter for the 10,000 cells for the treatment group and one minus the propensity score as the Poisson parameter for the

Figure 1
The True Propensity Score for Hypothetical Example 3 as a Function of *A* and *B*



10,000 cells for the control group. This sampling scheme generates (on average across simulated data sets) the equivalent of 10,000 sample members, assigned to the treatment instead of the control as a function of the probabilities plotted in Figure 1.[16]

Across the 50,000 simulated data sets, on average, 7,728 of the 10,000 possible combinations of values for both *A* and *B* had no individuals assigned to the treatment, and 4,813 had no individuals assigned to the control. No matter the actual realized pattern of sparseness for each simulated data set, all of the 50,000 data sets are afflicted, such that a perfect stratification on all values for the variables *A* and *B* would result in many strata within each data set for which only treatment or control cases are present.

To define treatment effects for each data set, two potential outcomes were defined as linear functions of individual values for A_i and B_i:

$$Y_i^1 = 102 + 6A_i + 4B_i + \varepsilon_i^1,$$
$$Y_i^0 = 100 + 3A_i + 2B_i + \varepsilon_i^0,$$

where both ε_i^1 and ε_i^0 are independent random draws from a normal distribution with expectation 0 and standard deviation of 5.[17] Then, as in equation (observed outcome definition), individuals from the treatment group were given an observed Y_i equal to their simulated Y_i^1, and individuals from the control group were given an observed Y_i equal to their simulated Y_i^0.

With this setup, the simulation makes available 50,000 data sets where the individual treatment effects can be calculated exactly since true values of Y_i^1 and Y_i^0 are available for all simulated individuals. Because the true average treatment effect, treatment effect for the treated, and treatment effect for the untreated are thus known for each simulated data set, these average effects can then serve as baselines against which alternative estimators that use data only on Y_i, D_i, A_i, and B_i can be compared.

The first row of Table 5 presents true Monte Carlo means and standard deviations of the three average treatments effects, calculated across the $50,000$ simulated data sets. The mean of the average treatment effect across data sets is 4.525, whereas the means of the average treatment effects for the treated and for the untreated are 4.892 and 4.395, respectively. Similar to hypothetical Example 1, this example represents a form of positive selection, where those who are most likely to be in the treatment group are also those most likely to benefit from the treatment. Accordingly, the treatment effect for the treated is larger than the treatment effect for the untreated.

Methods for treatment effect estimation. Rows 2 through 5 of Table 5 report means and standard deviations across the 50,000 data sets of three sets of regression estimates of the causal effect of D on Y. The first set is drawn from $50,000$ separate regressions of Y_i on D_i, resulting in parameter estimates exactly equivalent to what were defined in equation (6) as naive estimates (because, again, they do not use any information about the treatment assignment mechanism). The second and third sets of regression estimates incorporate least squares adjustments for A and B under two different specifications, linear and quadratic.

The last three rows of Table 5 present analogous results for three propensity score–based weighting estimators. For the estimates in the fifth

Table 5
Monte Carlo Means and Standard Deviations of Treatment Effects and Treatment Effect Estimates for Hypothetical Example 3

	Average Treatment Effect	Average Treatment Effect for the Treated	Average Treatment Effect for the Untreated
True treatment effects	4.525	4.892	4.395
	(.071)	(.139)	(.083)
Ordinary least squares regression estimators			
Regression of Y on D	5.388		
(i.e., the naive estimator)	(.121)		
Regression of Y on D, A, and B	4.753		
	(.117)		
Regression of Y on D, A, A^2, B, and B^2	4.739		
	(.118)		
Propensity score–based weighting estimators			
Misspecified propensity score estimates	4.456	4.913	4.293
	(.122)	(.119)	(.128)
Perfectly specified propensity score estimates	4.526	4.892	4.396
	(.120)	(.127)	(.125)
True propensity scores	4.527	4.892	4.396
	(.127)	(.127)	(.132)

row, it is (wrongly) assumed that the propensity score can be estimated consistently using a logit model with linear terms for A and B—that is, assuming that, for equation (17), a logit with $\mathbf{S}_i\boldsymbol{\phi}$, specified as $\alpha + \phi_A A_i + \phi_B B_i$, will yield consistent estimates of the propensity score surface plotted in Figure 1. After the logit model was estimated for each of the 50,000 data sets using the wrong specification, the estimated propensity score for each individual was then calculated:

$$\hat{p}_i = \frac{\exp(\hat{\alpha} + \hat{\phi}_A A_i + \hat{\phi}_B B_i)}{1 + \exp(\hat{\alpha} + \hat{\phi}_A A_i + \hat{\phi}_B B_i)} \tag{18}$$

along with the estimated odds of the propensity of being assigned to the treatment:

$$\hat{r}_i = \frac{\hat{p}_i}{1 - \hat{p}_i}, \tag{19}$$

where $\hat{p}_i$ is as constructed in equation (18).

To estimate the treatment effect for the treated, we then implemented a weighting estimator by calculating the average outcome for the treated and subtracting from this average outcome a counterfactual average outcome using weighted data on those from the control group:

$$\hat{\delta}_{TT,weight} \equiv \left\{ \frac{1}{n^1} \sum_{i:D_i=1} Y_i \right\} - \left\{ \frac{\sum_{i:D_i=0} \hat{r}_i Y_i}{\sum_{i:D_i=0} \hat{r}_i} \right\}, \tag{20}$$

where n^1 is the number of individuals in the treatment group, and $\hat{r}_i$ is the estimated odds of being in the treatment group instead of the control group, as constructed in equations (18) and (19). The weighting operation in the second term gives more weight to control group individuals equivalent to those in the treatment group (see Rosenbaum 1987, 2002). As we will describe later when discussing the connections between matching and regression, the weighting estimator in equation (20) can be written as a weighted ordinary least squares estimator.

To estimate the treatment effect for the untreated, we then implemented a weighting estimator that is the mirror image of the one in equation (20):

$$\hat{\delta}_{TUT,weight} \equiv \left\{ \frac{\sum_{i:D_i=1} Y_i/\hat{r}_i}{\sum_{i:D_i=1} n^1/\hat{r}_i} \right\} - \left\{ \frac{1}{n^0} \sum_{i:D_i=0} Y_i \right\}, \tag{21}$$

where n^0 is the number of individuals in the control group. Finally, the corresponding estimator of the unconditional average treatment effect is

$$\hat{\delta}_{ATE,weight} \equiv \left\{ \frac{1}{n} \sum_i D_i \right\} \left\{ \hat{\delta}_{TT,weight} \right\} + \left\{ 1 - \left[\frac{1}{n} \sum_i D_i \right] \right\} \left\{ \hat{\delta}_{TUT,weight} \right\}, \tag{22}$$

where $\hat{\delta}_{TT,weight}$ and $\hat{\delta}_{TUT,weight}$ are as defined in equations (20) and (21), respectively. Accordingly, an average treatment effect estimate is simply a weighted average of the two conditional average treatment effect estimates.

The same basic weighting scheme is implemented for the sixth row of Table 5, but the estimated propensity score used to define the estimated odds of treatment, $\hat{r}_i$, is instead based on results from a flawlessly estimated propensity score equation (i.e., one that uses the same specification that was fed to the random generator that assigned individuals to the treatment; see note 17 for the specification). Finally, for the last row of Table 5, the same weighting scheme is implemented, but in this case, the estimated odds of

treatment, $\hat{r}_i$, are replaced by the true odds of treatment, r_i, as calculated with reference to the exact function that generated the propensity score for Figure 1.

Monte Carlo results. As reported in the second through fourth rows of Table 5, all three regression-based estimators yield biased estimates of the average treatment effect (which are, on average, too large).[18] As reported in the fifth row of Table 5, the weighting estimator based on the misspecified logit yields estimates that are closer on average than the regression-based estimators for the average treatment effect. This difference is somewhat artificial since, in general, such a difference would depend on the relative misspecification of the propensity score estimating equation, the specification of the alternative regression equation, and the distributions of the potential outcomes.

The sixth row of Table 5 presents analogous estimates with flawlessly estimated propensity scores. These estimates are asymptotically unbiased and consistent for the average treatment effect, the treatment effect for the treated, and the treatment effect for the untreated. Finally, for the last row, the weighting estimates use the true propensity scores and are also asymptotically unbiased and consistent (but, as shown by Rosenbaum 1987, more variable than those based on the flawlessly estimated propensity score; see also Hahn 1998; Hirano, Imbens, and Ridder 2003; Rosenbaum 2002). The last two lines are thus the most important to note, as they demonstrate the most important claim of the literature: If one can obtain unbiased and consistent estimates of the true propensity scores, one can solve entirely the problems created by sparseness of data.

This example shows the potential power of propensity score–based modeling. If treatment assignment can be modeled perfectly, one can solve the sparseness problems that afflict finite data sets, at least insofar as offering estimates that are unbiased and consistent. On the other hand, this simulation also develops an important qualification of this potential power. Without a perfect specification of the propensity score estimating equation, one cannot rest assured that unbiased and consistent estimates can be obtained. Since propensity scores achieve their success by ''undoing'' the treatment assignment patterns, analogously to weighting a stratified sample, systematically incorrect estimated propensity scores can generate systematically incorrect weighting schemes that yield biased and inconsistent estimates of treatment effects. There is also the larger issue of whether the challenges of causal inference can be reduced to mere concerns about conditionally random sparseness, and this will depend entirely

on whether one is justified in imposing assumptions on the potential outcomes and treatment assignment process, as outlined earlier.

Given the description of matching estimators offered in the introduction (i.e., algorithms for mechanically identifying matched sets of equivalent treatment and control cases), in what sense are the individual-level weighting estimators of hypothetical Example 3 equivalent to matching estimators?

As emphasized earlier for hypothetical Examples 1 and 2, stratification estimators have a straightforward connection to matching. The strata that are formed represent matched sets, and a weighting procedure is then used to average stratified treatment effect estimates to obtain the average treatment effects of interest. The propensity score weighting estimators, however, have a less straightforward connection. Here, the data are, in effect, stratified coarsely by the estimation of the propensity score (i.e., since all individuals in the same strata, as defined by the stratifying variables in **S**, are given the same estimated propensity scores), and then the weighting is performed directly across individuals instead of across the strata. This type of individual-level weighting is made necessary because of sparseness (since some of the fine strata for which propensity scores are estimated necessarily contain only treatment or control cases, thereby preventing the direct calculation of stratified treatment effect estimates). Nonetheless, the same principle of balancing holds: Individuals are weighted within defined strata to ensure that the distribution of **S** is the same within the treatment and control cases that are then used to estimate the treatment effects.

In the opposite direction, it is important to recognize that the algorithmic matching estimators that we summarize in the next section can be considered weighting estimators. As we show later, these data analysis procedures warrant causal inference by achieving an "as if" stratification of the data that results in a balanced distribution of covariates across matched treatment and control cases. Thus, although it is sometimes easier to represent matching estimators as algorithmic data analysis procedures that mechanically match seemingly equivalent cases to each other, it is best to understand matching as a method to weight the data in order to warrant causal inference by balancing **S** across the treatment and control cases.

Matching as Data Analysis Algorithms

Algorithmic matching estimators differ primarily in (1) the number of matched cases designated for each to-be-matched target case and (2) how multiple matched cases are weighted if more than one is used for each

target case. In this section, we describe the four main types of matching estimators.

Heckman, Ichimura, and Todd (1997, 1998) and Smith and Todd (2005) outline a general framework for representing alternative matching estimators, and we follow their lead. Using our notation, all matching estimators of the treatment effect for the treated would be defined in this framework as

$$\hat{\delta}_{TT,match} = \frac{1}{n^1}\sum_i \left\{ (Y_i|D_i = 1) - \sum_j \omega_{i,j}(Y_j|D_j = 0) \right\}, \tag{23}$$

where n^1 is the number of treatment cases, i is the index over treatment cases, j is the index over control cases, and $\omega_{i,j}$ represents a set of scaled weights that measure the distance between each control case and the target treatment case. In equation (23), the weights are entirely unspecified.

Alternative matching estimators of the treatment effect for the treated can be represented as different procedures for deriving the weights represented by $\omega_{i,j}$. As we will describe next, the weights can take on many values—indeed, as many n^1-by-n^0 different values—since alternative weights can be used when constructing the counterfactual value for each target treatment case. The difference in the propensity score is the most common distance measure used to construct weights. Other measures of distance are available, including the estimated odds of the propensity score, the difference in the index of the estimated logit, and the Mahalanobis metric.[19]

Before describing the four main types of matching algorithms, we note two important points. First, for simplicity of presentation, in this section, we focus on matching estimators of the treatment effect for the treated. Each of the following matching algorithms could be used in reverse, instead focusing on matching treatment cases to control cases to construct an estimate of the treatment effect for the untreated. We mention this, in part, because it is sometimes implied in the applied literature that the matching techniques that we are about to summarize are only useful for estimating the treatment effect for the treated. This is false. If (1) all variables in **S** are known and observed, such that a perfect stratification of the data could be formed with a suitably large data set because both Assumptions 1-S and 2-S in equations (13) and (14) are valid, and (2) the ranges of all variables in **S** are the same for both treatment and control cases, then simple variants of the matching estimators that we present in this section can be formed that are consistent for the treatment effect among the

treated, the treatment effect among the untreated, and the average treatment effect.

Moreover, to consistently estimate the treatment effect for the treated, one does not need to assume full ignorability of treatment assignment or that both Assumptions 1-S and 2-S in equations (13) and (14) are valid. Instead, only Assumption 2-S (i.e., $E[Y_i^0|D_i=1,\mathbf{S}_i]=E[Y_i^0|D_i=0,\mathbf{S}_i]$) must hold.[20] In other words, to estimate the average treatment effect among the treated, it is sufficient to assume that, conditional on **S**, the average level of the outcome under the control for those in the treatment is equal, on average, to the average level of the outcome under the control for those in the control group.[21] This assumption is still rather stringent, as it asserts that those in the control group do not disproportionately gain from the control more than would those in the treatment group if they were instead in the control group. But it is surely weaker than having to assert Assumptions 1-S and 2-S together (which is again weaker than having to assert an assumption of strong ignorability of treatment assignment).

Second, as we show in a later section, the matching algorithms we summarize next are data analysis procedures that can be used more generally even when some of the variables in **S** are unobserved. The matching estimators may still be useful, as argued by Rosenbaum (2002), as a set of techniques that generate a provisional set of causal effect estimates that can then be subjected to further analysis. We discuss what sorts of further analysis have been proposed in the section that follows.

Basic Variants of Matching Algorithms

Exact Matching

For the treatment effect for the treated, exact matching constructs the counterfactual for each treatment case using the control cases with identical values on the variables in **S**. In the notation of equation (23), exact matching uses weights equal to $1/k$ for matched control cases, where k is the number of matches selected for each target treatment case. Weights of 0 are given to all unmatched control cases. If only one match is chosen randomly from among possible exact matches, then $\omega_{i,j}$ is set to 1 for the randomly selected match (from all available exact matches) and 0 for all other control cases. Exact matching may be combined with any of the matching methods described below.

Nearest Neighbor Matching

For the treatment effect for the treated, nearest neighbor matching constructs the counterfactual for each treatment case using the control cases that are closest to the treatment case on a unidimensional measure constructed from the variables in **S**, usually an estimated propensity score but sometimes variants of propensity scores (see Althauser and Rubin 1970; Rubin 1973a, 1973b, 1976a, 1976b, 1980; Rosenbaum and Rubin 1983a, 1985a, 1985b). The traditional algorithm randomly orders the treatment cases and then selects for each treatment case the control case with the smallest distance. The algorithm can be run with or without replacement. With replacement, a control case is returned to the pool after a match and can be matched later to another treatment case. Without replacement, a control case is taken out of the pool once it is matched.[22]

If only one nearest neighbor is selected for each treatment case, then $\omega_{i,j}$ is set equal to 1 for the matched control case and 0 for all other control cases. One can also match multiple nearest neighbors to each target treatment case, in which case $\omega_{i,j}$ is set equal to $1/k_i$ for the matched nearest neighbors, where k_i is the number of matches selected for each target treatment case i. Matching more control cases to each treatment case results in lower expected variance of the treatment effect estimate but also raises the possibility of greater bias since the probability of making more poor matches increases.

A danger with nearest neighbor matching is that it may result in some very poor matches for treatment cases. A version of nearest neighbor matching, know as caliper matching, is designed to remedy this drawback by restricting matches to some maximum distance. With this type of matching, some treatment cases may not receive matches, and thus the effect estimate will apply only to the subset of the treatment cases matched (even if ignorability holds and there is simply sparseness in the data).[23]

Interval Matching

Interval matching (also referred to as subclassification and stratification matching) sorts the treatment and control cases into segments of a unidimensional metric, usually the estimated propensity score, and then calculates the treatment effect within these intervals (see Cochran 1968; Rosenbaum and Rubin 1983a, 1984; Rubin 1977). For each interval, a variant of the matching estimator in equation (23) is estimated separately,

with $\omega_{i,j}$ chosen to give the same amount of weight to the treatment cases and control cases within each interval. The average treatment effect for the treated is then calculated as the mean of the interval-specific treatment effects, weighted by the number of treatment cases in each interval. This method is nearly indistinguishable from nearest neighbor caliper matching with replacement when each of the intervals includes exactly one treatment case.

Kernel Matching

Developed by Heckman et al. (Heckman, Ichimura, and Todd 1997, 1998; Heckman, Ichimura, Smith, and Todd 1998), kernel matching constructs the counterfactual for each treatment case, using all control cases, but weights each control case based on its distance from the treatment case. The weights represented by $\omega_{i,j}$ in equation (23) are calculated using a kernel function, $G(.)$, that transforms the distance between the selected target treatment case and all control cases in the study. When using the estimated propensity score to measure the distance, kernel-matching estimators define the weight as

$$\omega_{ij} = \frac{G\left(\frac{P(\mathbf{S}_j) - P(\mathbf{S}_i)}{a_n}\right)}{\sum_j G\left(\frac{P(\mathbf{S}_j) - P(\mathbf{S}_i)}{a_n}\right)}, \tag{24}$$

where a_n is a bandwidth parameter that scales the difference in the estimated propensity scores based on the sample size, and $P(.)$ is the estimated propensity score as a function of its argument.[24] The numerator of this expression yields a transformed distance between each control case and the target treatment case. The denominator is a scaling factor equal to the sum of all the transformed distances across control cases, which is needed so that the sum of $\omega_{i,j}$ is equal to 1 across all control cases when matched to each target treatment case.

Although kernel-matching estimators appear quite complex, they are a natural extension of interval and nearest neighbor matching: All control cases are matched to each treatment case but weighted so that those closest to the treatment case are given the greatest weight. Smith and Todd (2005) offer an excellent intuitive discussion of kernel matching along with generalizations to local linear matching (Heckman, Ichimura, and Todd 1997, 1998; Heckman, Ichimura, Smith, and Todd 1998) and local quadratic matching (Ham, Li, and Reagan 2003).

Which of These Basic Matching Algorithms Works Best?

There is very little specific guidance in the literature on which of these matching algorithms works best, and the answer very likely depends on the substantive application. Smith and Todd (2005) and Heckman et al. (Heckman, Ichimura, and Todd 1997, 1998; Heckman, Ichimura, Smith, and Todd 1998) have experimental data against which matching estimators can be compared, and they argue for the advantages of kernel matching (and a particular form of robust kernel matching). To the extent that a general answer to this question can be offered, we would suggest that nearest neighbor caliper matching with replacement, interval matching, and kernel matching are all closely related and should be preferred to nearest neighbor matching without replacement. If the point of a matching estimator is to minimize bias by comparing target cases to similar matched cases, then methods that make it impossible to generate poor matches should be preferred.[25] It is also sometimes advisable to combine matching with regression adjustment if there is a question as to whether balance has been achieved (see Rubin and Thomas 2000). Matching on both the propensity score and the Mahalanobis metric has also been recommended for achieving balance on higher order moments (see Rosenbaum and Rubin 1985a, 1985b; Diamond and Sekhon 2005).[26]

Since there is no clear guidance on which of these matching estimators is "best," we constructed a fourth hypothetical example to give a sense of how often alternative matching estimators yield appreciably similar estimates. We also develop this example so that it can serve as a bridge to the section that follows, where the substantial additional challenges of real-world applications are discussed.

Hypothetical Example 4

For this example, we use simulated data, where we defined the potential outcomes and treatment assignment patterns so that we can explore the relative performance of alternative matching and regression estimators. The former are estimated under alternative scenarios with two different specifications of the propensity score estimating equation. Unlike hypothetical Example 3, we do not repeat the simulation for multiple samples but confine ourselves to results on a single sample, as would be typical of any real-world application.

Generation of the data set. The data set that we constructed mimics the data set from the National Education Longitudinal Study analyzed by Morgan (2001). For that application, Morgan used regression and matching estimators to estimate the effect of Catholic schooling on the achievement of high school students in the United States. For our simulation, we generated a data set of 10,000 individuals with values for 13 baseline variables that resemble closely the joint distribution of the similar variables in Morgan. The variables include dummies for race, region, urbanicity, have own bedroom, and have two parents, along with an ordinal variable for number of siblings and a continuous variable for socioeconomic status. Then, we created an entirely hypothetical cognitive skill variable, assumed to reflect innate and acquired skills in unknown proportions.[27]

We then defined potential outcomes for all $10,000$ individuals, assuming that the observed outcome of interest is a standardized test taken at the end of high school. For the potential outcome under the control (i.e., a public school education), we generated "what-if" test scores form a normal distribution, with an expectation as a function of race, region, urbanicity, number of siblings, socioeconomic status, family structure, and cognitive skills. We then assumed that the "what-if" test scores under the treatment (i.e., a Catholic school education) would be equal to the outcome under the control plus a boosted outcome under the treatment that is function of race, region, and cognitive skills (under the assumption, based on the dominant position in the extant literature, that black and Hispanic respondents from the north, as well as all of those with high cognitive skills, are disproportionately likely to benefit from Catholic schooling).

We then defined the probability of attending a Catholic school using a logit with 26 parameters, based on a specification from Morgan (2001) along with an assumed self-selection dynamic where individuals are slightly more likely to select the treatment as a function of the relative size of their individual-level treatment effect.[28] This last component of the logit creates a nearly insurmountable challenge since in any particular application, one would not have such a variable with which to estimate a propensity score. That, however, is our point in including this term, as individuals are thought, in many real-world applications, to be selecting from among alternative treatments based on accurate expectations, unavailable as measures to researchers, of their likely gains from alternative treatment regimes. The probabilities defined by the logit were then passed to a binomial distribution, which resulted in 986 of the $10,000$ simulated students attending Catholic schools. Finally, observed outcomes were assigned according to treatment status.

With the sample divided into the treatment group and the control group, we calculated from the prespecified potential outcomes the true baseline average treatment effects. The treatment effect for the treated is 6.96 in the simulated data, while the treatment effect for the untreated is 5.9. In combination, the average treatment effect is then 6.0.

Methods for treatment effect estimation. In Table 6, we offer 12 separate types of matching estimates, as well as an additional 5 that incorporate supplemental regression adjustment. These are based on routines written for STATA by three sets of authors: Abadie et al. (2001), Becker and Ichino (2002), and Leuven and Sianesi (2003).[29] We estimate all matching estimators under two basic scenarios. First, we offer a set of estimates based on poorly estimated propensity scores, derived from an estimating equation from which we omitted nine interaction terms along with the cognitive skill variable. The last specification error is particularly important, as the cognitive skill variable has a correlation of 0.401 with the outcome and 0.110 with the treatment in the simulated data. For the second scenario, we included the cognitive skill variable and the nine interaction terms. Both scenarios lack an adjustment for the self-selection dynamic, in which individuals select into the treatment partly as a function of their expected treatment effect.

Regarding the specific settings for the alternative matching estimators, which are listed in the row headings of Table 6, the interval matching algorithm began with five blocks and subdivided blocks until each block achieved balance on the estimated propensity score across treatment and control cases. Nearest neighbor matching with replacement was implemented with and without a caliper of 0.001, in both one and five nearest neighbor variants. Radius matching was implemented using a radius of 0.001. For the kernel-matching estimators, we used two types of kernels—Epanechnikov and Gaussian—and the default bandwidth of 0.06 for both pieces of software. For the local linear matching estimator, we used the Epanechnikov kernel with the default bandwidth of 0.08.

For comparison, we offer OLS regression estimates of the treatment effect under two analogous scenarios (i.e., including the same variables for the propensity score estimating equation directly in the regression equation). We present regression estimates in two variants: (1) without regard to the distributions of the variables and (2) based on a subsample restricted to the region of common support (as defined by the propensity score estimated from the covariates used for the respective scenario).

Finally, we provide five examples of matching combined with regression adjustment. Interval matching with regression adjustment calculates

Table 6
Matching and Regression Estimates for the Simulated Effect of Catholic Schooling on Achievement, as Specified in Hypothetical Example 4

	Poorly Specified Propensity Score Estimating Equation		Well-Specified Propensity Score Estimating Equation	
	TT Estimate	Bias	TT Estimate	Bias
Matching				
Interval with variable blocks (B&I)	7.93	0.97	6.73	–0.23
One nearest neighbor with caliper = 0.001 (L&S)	8.16	1.20	6.69	–0.27
One nearest neighbor without caliper (Abadie)	7.90	0.94	6.62	–0.34
Five nearest neighbors with caliper = 0.001 (L&S)	7.97	1.01	7.04	0.08
Five nearest neighbors without caliper (Abadie)	7.85	0.89	7.15	0.19
Radius with radius = 0.001 (L&S)	8.02	1.06	6.90	–0.06
Radius with radius = 0.001 (B&I)	8.13	1.17	7.29	0.33
Kernel with Epanechnikov kernel (L&S)	7.97	1.01	6.96	0.00
Kernel with Epanechnikov kernel (B&I)	7.89	0.93	6.86	–0.10
Kernel with Gaussian kernel (L&S)	8.09	1.13	7.18	0.22
Kernel with Gaussian kernel (B&I)	7.97	1.01	7.03	0.09
Local linear with Epanechnikov kernel (L&S)	7.91	0.95	6.84	–0.12
Ordinary least squares regression				
Not restricted to region of common support	7.79	0.83	6.81	–0.15
Restricted to region of common support	7.88	0.92	6.80	–0.16

(continued)

Table 6 (continued)

	Poorly Specified Propensity Score Estimating Equation		Well-Specified Propensity Score Estimating Equation	
	TT Estimate	Bias	TT Estimate	Bias
Matching with regression				
Interval with variable blocks & regular adjustment (B&I)	7.95	0.99	6.70	–0.26
One nearest neighbor with caliper = 0.001 & regular adjustment (L&S)	8.05	1.09	7.15	0.19
One nearest neighbor without caliper & with regular adjustment (Abadie)	7.78	0.82	6.88	–0.08
Five nearest neighbors with caliper = 0.001 & regular adjustment (L&S)	7.92	0.96	7.17	0.21
Five nearest neighbors without caliper & with regular adjustment (Abadie)	7.82	0.86	7.20	0.24

Note: B&I denotes the Becker and Ichino software. L&S denotes the Leuven and Sianesi software. Abadie denotes the Abadie et al. software. TT = treatment effect on the treated.

the treatment effect within blocks after adjusting for the same covariates included in the propensity score estimating equation for the particular scenario, averaging over blocks to produce an overall treatment effect estimate. With nearest neighbor matching, regression adjustment is accomplished by regressing the outcome on the treatment and covariates using the matched sample, with appropriate weights for duplicated observations in the matched control group and for multiple neighbor matching.

Results. We estimated treatment effects under the assumption that self-selection on the individual Catholic school effect is present and yet cannot be adjusted for using a statistical model without a measure of individuals' expectations. Thus, we operate under the assumption that only the treatment effect for the treated has any chance of being estimated consistently, as in the study by Morgan (2001) on which this example is based. We therefore compare all estimates to the true simulated treatment effect for the treated, identified earlier as 6.96.

Estimates using the poorly estimated propensity scores are reported in the first column of Table 6, along with the implied bias as an estimate of the treatment effect for the treated in the second column (i.e., the matching estimate minus 6.96). As expected, all estimates have a substantial positive bias. Most of the positive bias results from the mistaken exclusion of the cognitive skill variable from the propensity score estimating equation.

Matching estimates using the well-estimated propensity scores are reported in the third column of Table 6, along with the expected bias in the fourth column. On the whole, these estimates are considerably better. Having the correct specification reduces the bias in those estimates with the largest bias from column 3, and on average, all estimates oscillate around the true treatment effect for the treated of 6.96.[30]

For comparison, we then provide analogous regression estimates in the second panel of the table. In some cases, these estimates outperform some of the matching estimates. In fairness to the matching estimates, however, it should be pointed out that the data analyzed for this example are well suited to regression because the assumed functional form of each potential outcome variable is linear and hence relatively simple. Although we believe that this is reasonable for the simulated application, there are surely scenarios in which matching can be shown to clearly outperform regression because of nonlinearities.

The final panel of the table presents estimates from matching combined with regression adjustment. In several cases, regression adjustment provides

a slight improvement over the analogous matching estimator implemented without regression adjustment. But this is not true in all cases.

We have demonstrated three basic points with this example. First, looking across the rows of Table 6, it is clear that matching estimators and different software routines yield different treatment effect estimates (even ones that are thought to be mathematically equivalent). Thus, at least for the near future, it will be crucial for researchers to examine multiple estimates of the same treatment effect across estimators and software packages. We found the lack of consistency across seemingly equivalent estimators from alternative software routines to be somewhat troubling, but we assume that this unexpected variation will dissipate with improvements in the software.

Second, matching estimators cannot compensate for an unobserved covariate in **S**, which leads to comparisons of treatment and control cases that are not identical in all relevant aspects other than treatment status. The absence of the cognitive skill variable for the poorly estimated propensity scores invalidates both Assumptions 1-S and 2-S. The matching routines still balance the variables included in the propensity score estimating equation, but the resulting matching estimates remain biased and inconsistent.

Third, the sort of self-selection dynamic built into this example—where individuals choose Catholic schooling as a function of their expected gains from Catholic schooling—makes estimation of both the average treatment effect among the untreated and the average treatment effect impossible. Even if all variables in **S** are observed (i.e., including cognitive skill in this example), only the average treatment effect among the treated can be estimated consistently because Assumption 1-S cannot be maintained.[31]

Unfortunately, violation of the assumption of ignorable treatment assignment (and of both Assumptions 1-S and 2-S) is the scenario in which most analysts will find themselves, and this is the scenario to which we turn in the next section. Before discussing what can be done when ignorability of any form cannot be assumed, we first close the discussion on which types of matching may work best.

Matching Algorithms That Seek Optimal Balance

For Example 4, we judged the quality of matching algorithms by examining the distance between the treatment effect estimates that we obtained and the true treatment effects that we stipulated in constructing our hypothetical data. Because we only generated one sample, these differences are not necessarily a very good guide to practice, even though our

main goal of the example was to show that alternative matching estimators generally yield different results, and in the absence of ignorability, none of these may be correct. That example aside, it is generally recognized that the best matching algorithms are those that optimize balance in the data being analyzed. Building on this consensus, a broader set of matching algorithms is currently in development, which grows out of the optimal matching proposals attributed to Rosenbaum (1989).

Matching is generally judged to be successful if, for both the treatment and matched control groups, the distribution of the matching variables is the same. When this result is achieved, the data are said to be balanced, as noted earlier. As shown by Rosenbaum and Rubin (1984), balance can be assessed quickly using paired *t* tests for differences in the means of the matching variables across matched treatment and control cases. But, to achieve full balance, the entire joint distribution of the matching variables must be the same, with all observed differences small enough to be attributable to random variation. To meet this standard, one must evaluate the equivalence of full joint distributions, and more complicated tests are required (such as nonparametric Kolmogorov-Smirnov tests; see Abadie 2002 and Diamond and Sekhon 2005).

If the covariates are not balanced, one can change the estimation model for the propensity score, for example, by adding interaction terms, quadratic terms, or other higher order terms. Or, one can match on the Mahalanobis metric in addition to the propensity score, perhaps nesting one set of matching strategies within another. This respecification is not considered data mining because it does not involve examining the effect estimate. But it can be labor intensive, and there is no guarantee that one will find the best possible balance by simply reestimating the sorts of matching algorithms introduced earlier or combining them in novel ways.

For this reason, two more general forms of matching have been proposed, each of which is now fairly well developed (but not easy to implement in standard data analysis packages commonly used in sociology). Rosenbaum (2002, chap. 10) reports on recent results for full optimal matching algorithms that he has achieved with colleagues since Rosenbaum (1989). His algorithms seek to optimize balance and efficiency of estimation by searching through all possible matches that could be made, after stipulating the minimum and maximum number of matches for matched sets of treatment and control cases. Although full optimal matching algorithms vary (see also Hansen 2004a), they are based on the idea of minimizing the average distance between the estimated propensity scores

among matched cases. If the estimated propensity scores are correct, then this minimization problem should balance **S**.

Diamond and Sekhon (2005) propose a general multivariate matching method that uses a genetic algorithm to search for the match that achieves the best possible balance. The quality of balance is specified as a standard set of *t* tests of differences of means but also bootstrapped Kolmogorov-Smirnov tests for the full distributions of the matching variables. Although their algorithm can be used to carry out matching after the estimation of a propensity score, their technique is more general and can almost entirely remove the analyst from having to make any specification choices other than designating the matching variables that one wishes to balance. Diamond and Sekhon show that their matching algorithms provide superior balance in both Monte Carlo simulations and a test with genuine data.

Although there is good reason to expect that these types of matching algorithms can outperform the nearest neighbor, interval, and kernel-matching algorithms by the criteria of balance, they are considerably more difficult to implement in practice. With software developments under way, these disadvantages will be eliminated.

Matching When Treatment Assignment Is Nonignorable

What if neither Assumption 1-S nor Assumption 2-S is viable because one only observes a subset of the variables in **S**, which we will now denote by **X**? One can still match on **X** using the techniques just summarized, as we did for the first column of Table 6 in hypothetical Example 4.

When in this position, one should concentrate on estimating only one type of treatment effect (usually the treatment effect for the treated, although perhaps the unconditional average treatment effect). Because a crucial step must be added to the project—assessing the level of bias that may arise from possible nonignorability of treatment—focusing on a very specific treatment effect of primary interest helps to ground a discussion of an estimate's limitations. Then, after using one of the matching estimators of the last section, one should use the data to minimize bias in the estimates and, if possible, proceed thereafter to a sensitivity analysis. We discuss the possibilities for these steps in the order that analysts usually carry them out.

Covariance adjustment can be incorporated easily into matching estimators. Two alternative but similar methods exist. Rubin and Thomas (2000; see also 1996) propose a method that can be used in conjunction

with nearest neighbor and interval matching. One simply estimates a regression model on the data set created by the matching procedure, perhaps reusing some or all of the variables in **X**, in hopes of relieving unknown consequences of any slight misspecification of the propensity score estimating equation. The covariates are simply included in the regression model alongside D, possibly with fixed effects for alternative strata if multiple cases have been matched to each target case. These methods are implemented in hypothetical Example 4 and need not be used only in cases where only a subset of **S** is observed. In fact, Robins and his colleagues (see van der Laan and Robins 2003 for citations) have argued in a series of papers that, in general, one should always offer such "doubly robust" estimates of causal effects, in hopes that misspecifications of the propensity score estimating equation and the final regression equation will neutralize each other.

Heckman et al. (Heckman, Ichimura, and Todd 1997, 1998; Heckman, Ichimura, Smith, and Todd 1998) propose a slightly different procedure. First, one regresses Y on covariates for those in the control group, saving the regression estimates in a vector β_c. Then, one creates predicted values for all individuals using the variables of particular interest by applying the estimated regression parameters to both the treatment and control cases. Finally, if estimating the treatment effect for the treated, one then offers matching estimates based on equation (23) using the residuals in place of the outcomes.

Abadie and Imbens (2004) show that failure to use a regression adjustment procedure in tandem with a matching algorithm can lead to bias in finite samples in analyses in which **S** contains more than one continuous variable. The amount of potential bias increases with the number of variables in the assignment equation. They recommend a simple linear regression adjustment, offering STATA and MATLAB programs that implement nearest neighbor matching along with the bias correction (see Abadie et al. 2001). We implemented these estimates for the last panel of Table 6, and in one of the two instances, the regression adjustments reduced the bias of the estimate.

Although these adjustment procedures may help to refine the balance of **X** across treatment and control cases, they do not help to address the problem of unobservable variables in **S**. These problems can be quite serious if the unobserved variables are fairly subtle, such as a differential latent growth rate for the outcome that is correlated with treatment assignment/selection. In such cases, the options are quite limited for using the data to diagnose and then correct bias in one's estimates.

If longitudinal data are available, one can incorporate a difference-in-difference adjustment into any of the matching estimators discussed earlier (see Smith and Todd 2005). For example, when data on the outcome prior to the treatment are available for both the treatment and control cases, one can substitute into equation (23) the difference between the posttreatment outcome and the pretreatment outcome for the posttreatment outcome. The difference-in-difference matching estimator attempts to account for all time-constant covariates and is analogous to adding individual fixed effects to a regression model. An alternative, as in Dehejia and Wahba (1999), is to include the pretreatment outcome in the regression equations estimated with the data set constructed by the matching procedure. In evaluations of matching estimates of the treatment effect of training programs, Heckman et al. (1997) and Smith and Todd (2005) find that a difference-in-difference local linear matching estimator performed well, coming closest to replicating the experimental estimates of the effect of the Job Training Partnership Act (JTPA) and National Supported Work (NSW) programs. Whether this optimal performance is a reasonable guide for other applications remains to be determined.

Finally, one can perform a sensitivity analysis and/or use the extant literature to discuss the heterogeneity that may lurk beneath the matching estimate. Harding (2003) and DiPrete and Gangl (2004), for example, draw on the tradition of Rosenbaum (1991, 1992; see also Rosenbaum and Rubin 1983b) to assess the strength of the relationship that an unobserved variable would have to have with a treatment and an outcome variable to challenge a causal inference. Morgan (2001) analyzes variation in the treatment effect estimate across quintiles of the estimated propensity score, offering alternative interpretations of variation in treatment effect estimates based on competing positions in the relevant applied literature about the nature of some crucial unobserved variables. Rosenbaum (2002) devotes a large portion of his excellent book on observational data analysis to strategies for performing sensitivity analysis to determine the potential impact of hidden bias on one's conclusions.

Remaining Practical Issues in Matching Analysis

In this section, we discuss the remaining practical issues that analysts who consider using matching estimators must confront. First, we discuss the relationship between matching estimators and more standard regression approaches that dominate empirical research in sociology. We show how

some matching estimators can be written (and hence understood) as well-specified regression models. Following directly on this discussion—which also shows how regression, nonetheless, can make it all too easy to compare incomparable individuals—we then discuss the practical issue of how to empirically identify the common support of the matching variables. Finally, we discuss what is known about the sampling variance of alternative matching estimators, and we give a guide to usage of the standard errors provided by existing software.

Matching and Regression

Our presentation aside, we must note that matching and regression are closely related methods, and each can be seen as a variant of the other. Consider how the matching estimates in our hypothetical Examples 1 and 3 could have been generated via standard regression routines.

For hypothetical Example 1, presented in Tables 1 and 2, an analyst could specify S as two dummy variables and D as one dummy variable. If all two-way interactions between S and D are then included in a regression model predicting the outcome, then one has enacted the same perfect stratification of the data by fitting a saturated model to the cells of the first panel of Table 2. Accordingly, if one obtains the marginal distribution of S and the joint distribution of S given D, then one can properly average the coefficient contrasts across the relevant distributions of S to obtain consistent estimates of the average treatment effect, the treatment effect among the treated, and the treatment effect among the untreated.

For hypothetical Example 3, presented in Table 5, the three propensity score weighting estimates in equations (20) through (22) could be specified as three weighted ordinary least squares regression models. In fact, if one defines a weighting variable appropriately, then any standard software package that estimates weighted regression can be used.

To see how to do this, note first that the naive estimator in equation (6) can be written as an OLS estimator, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, where (1) $\mathbf{X}$ is an n-by-2 matrix that contains a vector of 1s in its first column and a vector of the values of D for each individual in its second column, and (2) $\mathbf{y}$ is an n-by-1-column vector containing values of Y for each individual. To estimate each of the propensity score weighting estimators in equations (20) through (22), one simply estimates a weighted ordinary least squares estimator, $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$, with an appropriately chosen weight matrix $\mathbf{W}$. For $\hat{\delta}_{TT,weight}$ in equation (20), one specifies $\mathbf{W}$ as an n-by-n diagonal matrix with 1 in the i-by-ith place for members of the treatment group and

$\hat{p}_i/(1-\hat{p}_i)$ in the i-by-ith place for members of the control group (where, as defined earlier for hypothetical Example 3, $\hat{p}_i$ is the estimated propensity score). For $\hat{\delta}_{TUT,weight}$ in equation (21), one specifies $\mathbf{W}$ as an n-by-n diagonal matrix with $(1-\hat{p}_i)/\hat{p}_i$ in the i-by-ith place for members of the treatment group and 1s in the i-by-ith place for members of the control group. Finally, for $\hat{\delta}_{ATE,weight}$ in equation (22), one specifies $\mathbf{W}$ as an n-by-n diagonal matrix with $1/\hat{p}_i$ in the i-by-ith place for members of the treatment group and $1/(1-\hat{p}_i)$ in the i-by-ith place for members of the control group.[32]

More generally, the relationship between matching and regression has been established in the recent literature. Most matching estimators can be rewritten as forms of nonparametric regression (see Hahn 1998; Heckman, Ichimura, and Todd 1998; Hirano et al. 2003; Imbens 2004), and ordinary least squares regression can be seen as a variance-weighted form of interval matching (see Angrist and Krueger 1999). Moreover, all average causal effect estimators can be interpreted as weighted averages of marginal treatment effects (see Heckman and Vytlacil 2004), whether generated by matching, regression, or local instrumental variable estimators.

Nevertheless, even though regression can be used as a technique to execute a stratification of the data, and weighted regression can be used to estimate propensity score weighting estimators, regression can also yield misleading results. If, for hypothetical Example 1, S were entered as a simple linear term interacted with D (or, instead, if S were entered as two dummy variables but not interacted with D), regression would yield coefficient contrasts that mask the underlying treatment effects.[33] In a sense, this problem is simply a matter of model misspecification. But, at a deeper level, it may be that regression as a method tends to encourage the analyst to oversimplify these important model specification issues. Consider hypothetical Example 2, depicted in Tables 3 and 4. If a saturated regression model is fit to the data, the lack of overlap in the distribution of S will be revealed to the analyst when the regression routine drops the coefficient for the zero cell. However, if a constrained version of the model were fit, such as if S were entered as a simple linear term interacted with D, the regression would yield seemingly reasonable coefficients.

All too often, regression modeling, at least as practiced in sociology, makes it too easy for an analyst to overlook fundamental mismatches between treatment and control cases. And, thus, one can obtain average treatment effect estimates with regression techniques even when no meaningful average treatment effect exists. Sensitivity to these possibilities has

led to a specialized set of techniques to focus the attention of the analyst on the importance of these concerns.

Assessing the Region of Common Support

In practice, there is often good reason to believe that some of the lack of observed overlap of **S** for the treatment and control cases may have emerged from systematic sources, often related to the choice behavior of individuals (see Heckman, Ichimura, Smith, and Todd 1998). In these situations, it is not a sparseness problem that must be corrected. Instead, a more fundamental mismatch between the observed treatment and control cases must be addressed, as in our earlier hypothetical Example 2. Treatment cases that have no possible counterpart among the controls are said to be "off the support" of **S** for the control cases and likewise for control cases that have no possible counterparts among the treatment cases.[34]

When in this situation, applied researchers who use matching techniques to estimate the treatment effect for the treated often estimate a narrower treatment effect. Using one of the variants of the matching estimators outlined earlier, analysis is confined only to treatment cases whose propensity scores fall between the minimum and maximum propensity scores in the control group. Resulting estimates are then interpreted as estimates of a narrower treatment effect: the common-support treatment effect for the treated (see Heckman, Ichimura, and Todd 1997, 1998).

The goal of these sorts of techniques is to exclude at the outset those treatment cases that are beyond the observed minima and maxima of the probability distributions of the variables in **S** among the control cases (and vice versa). Although using the propensity score to find the region of overlap may not capture all dimensions of the common support (as there may be interior spaces in the joint distribution defined by the variables in **S**), subsequent matching is then expected to finish the job.

Sometimes, matching on the region of common support helps to clarify and sharpen the contribution of a study. When estimating the average treatment effect for the treated, there may be little harm in throwing away control cases outside the region of common support if all treatment cases fall within the support of the control cases. And even if imposing the common support condition results in throwing away some of the treatment cases, this can be considered an important substantive finding, especially for interpreting the treatment effect estimate. In this case, the resulting estimate is the treatment effect for a subset of the treated only and, in particular, a treatment effect estimate that is informative only about those in

the treatment and control groups who are equivalent with respect to observed treatment selection variables. In some applications, this is precisely the estimate needed (e.g., when evaluating whether a program should be expanded in size to accommodate more treatment cases but without changing eligibility criteria).[35]

Coming to terms with these common support issues has become somewhat of a specialized art form within the empirical matching literature, and some guidance is available. Heckman, Ichimura, and Todd (1998; see also Smith and Todd 2005) recommend trimming the region of common support to eliminate cases in regions of the common support with extremely low density (and not just with respect to the propensity score but for the full distribution of **S**). This involves selecting a minimum density (labeled the "trimming level") that is greater than zero. Heckman and his colleagues have found that estimates are rather sensitive to the level of trimming in small samples, with greater bias when the trimming level is lower. However, increasing the trimming level excludes more treatment cases and results in higher variance. Such concerns with the consequences for the variance of estimates are hard to judge, for as we discuss next, much remains to be learned about the statistical properties of matching estimators.

The Expected Variance of Matching Estimates

After computing a matching estimate of some form, most researchers naturally desire a measure of its expected variability across samples of the same size from the same population, either to conduct hypothesis tests or to offer an informed posterior distribution for the causal effect that can guide subsequent research. We did not, however, report standard errors for the treatment effect estimates reported in Table 6 for hypothetical Example 4. Most of the available software routines provide such estimates.

For example, for the software of Abadie and his colleagues, the one and five nearest neighbor matching estimates of 7.90 and 7.85 in the first column of Table 6 have estimated standard errors of .671 and .527, respectively. These values can be directly compared to two other sets of estimates. For the OLS estimates of 7.79 and 7.88, which differ depending on whether the sample is restricted to the region of common support as measured by the distribution of the estimated propensity score, the estimated standard errors are .451 and .452, respectively. And for the comparable one and five nearest neighbor matching estimates with regression

adjustment of 7.78 and 7.82, the estimated standard errors are .642 and .509, respectively. The relative sizes of these standard error estimates across methods are broadly consistent with what one finds in the applied matching literature: Regression yields the smallest estimated standard errors, and regression adjustments to matching estimators reduce the standard errors of matching estimates alone.

Nonetheless, each of the software routines we used relies on a different methodology for calculating such estimates, and given their mismatch, we caution against too strong of a reliance on the standard error estimates produced by any one software routine at present. Much remains to worked out before commonly accepted standards for calculating standard errors are available. For now, our advice is to report standard errors for regression estimates and then to give a sense of the range of standard errors produced by alternative software for corresponding matching estimates.[36]

We recommend caution for the following reasons. In some simple cases, there is widespread agreement on how to properly estimate standard errors for matching estimators. For example, if a perfect stratification of the data can be found, the data can be analyzed as if they are a stratified random sample, with the treatment randomly assigned within each stratum. In this case, the variance estimates from stratified sampling carry over. But rarely is a perfect stratification available in practice without substantial sparseness in the data at hand. Once stratification is performed with reference to an estimated propensity score, the independence that is assumed within strata for standard error estimates from stratified sampling methodology is no longer present. And if one adopts a Bayesian perspective, the model uncertainty of the propensity score estimating equation must be represented in the posterior.[37]

Even so, there is now also widespread agreement that convergence results from nonparametric statistics can be used to justify standard error estimates for large samples. A variety of scholars have begun to work out alternative methods for calculating such asymptotic standard errors for matching estimators, after first rewriting matching estimators as forms of nonparametric regression (see Abadie and Imbens 2006; Heckman, Ichimura, and Todd 1998; Hahn 1998; Hirano et al. 2003; Imbens 2004). For these large-sample approaches, however, it is generally assumed that matching is performed directly with regard to the variables in **S**, and the standard errors are appropriate only for large samples in which sparseness is vanishing. Accordingly, the whole idea of using propensity scores to solve rampant sparseness problems is almost entirely dispensed with, and estimated propensity scores then serve merely to clean up whatever chance

variability in the distribution of **S** across treatment and control cases remains in a finite sample.

How frequently will either of these ways of computing standards errors fit the situations in which applied sociologists will find themselves? In general, the samples with which sociologists work are of moderate size. Moreover, sparseness is widespread, and propensity scores (or some other lower dimensional function in **S**) must be used to formulate matches. Abadie and Imbens (2006) show that one can use brute-force computational methods to estimate sample variances at points of the joint distribution of **S**. When combined with nonparametric estimates of propensity scores, one can obtain consistent estimates of all pieces of their proposed formulas for asymptotic standard errors. And yet, none of this work shows that the variance estimators that have been proposed remain good guides for the expected sampling variance of matching estimators under different amounts of misspecification of the propensity score estimating equation or when matching is attempted only with regard to the estimated propensity score rather than completely on the variables in **S**. Given that this literature is still developing, it seems prudent to report alternative standard errors from alternative software routines and to avoid drawing conclusions that depend on accepting any one particular method for calculating standard errors.

Conclusions: Strengths and Weaknesses of Matching Estimators

We conclude by discussing the strengths and weaknesses of matching as a method for causal inference from observational data. Some of the advantages of matching methods are not inherent or unique to matching itself but rather are the result of the analytical framework in which most matching analyses are conducted. Matching focuses attention on the heterogeneity of the causal effect. It forces the analyst to examine the alternative distributions of covariates across those exposed to different levels of the causal variable. The process of examining the region of common support helps the analyst to recognize which cases in the study are incomparable, such as which control cases should be ignored when estimating the treatment effect for the treated and which treatment cases may have no meaningful counterparts among the controls. Finally, matching helps to motivate more sophisticated discussions of the unobservables that may be correlated with the causal variable, and this is an advance over merely conceding that selection

bias may be present in some form and speculating on the sign of the bias. Thus, although matching does not solve all (or even very many) of the problems that prevent regression models from generating reliable estimates of causal effects, matching succeeds admirably in laying bare the particular problems of estimating causal effects and then motivating the future research that is needed to resolve causal controversies.

There are some specific advantages of matching. When matching is accompanied by explicit balance testing, it minimizes the need to make assumptions about functional form. If covariates are balanced after matching, one has not relied on the functional form assumptions of the propensity score model. Thus, matching may significantly outperform regression when the true functional form of a regression is nonlinear but a simple linear specification is used. In addition, for nontechnical audiences, matching is often a more intuitive method for dealing with covariates than regression adjustment. The idea that treatment and control groups have the same distributions of observed covariates is often easier to explain than how one ostensibly "controls for" covariates using regression.

Although these are the advantages of matching, it is important that we not oversell the potential power of the techniques. First, even though the extension of matching techniques to multivalued treatments has begun, readily available matching estimators can only be applied to treatments or causal exposures that are binary. Second, as we just discussed, our inability to estimate the variance of most matching estimators with commonly accepted methods is a genuine weakness (although it is reasonable to expect that this weakness can be overcome in the near future). Third, as hypothetical Example 4 showed, different matching estimators can lead to somewhat different estimates of causal effects, and as yet, there is little guidance as to which types of matching estimators work best for different types of applications.

Finally, we close by drawing attention to a common misunderstanding about matching estimators. In much of the applied literature on matching, the propensity score is presented as a single predictive dimension that can be used to balance the distribution of important covariates across treatment and control cases, thereby warranting causal inference. As we showed in hypothetical Example 4, perfect balance on important covariates does not necessarily warrant causal claims. If one does not know of variables that, in an infinite sample, would yield a perfect stratification, then simply predicting treatment status from the observed variables using a logit model and then matching on the estimated propensity score does not solve the causal inference problem. The estimated propensity scores will balance those variables

across the treatment and control cases. But the study will remain open to the sort of "hidden bias" explored by Rosenbaum (2002) but that is often labeled selection on the unobservables in the social sciences. Matching, like regression, is thus a statistical method for analyzing available data, which may have some advantages in some situations. But, in the end, matching cannot compensate for data insufficiency. Causal controversies are best resolved by collecting new and better data.

Notes

1. A virtue of matching, as developed in this tradition, is cost-effectiveness for prospective studies. If the goal of a study is to measure the evolution of a causal effect over time by measuring symptoms at several points in the future, then discarding nontreated cases unlike any treated cases can cut expenses without substantially affecting the quality of causal inferences that a study can yield.

2. We adopt this general setup for expository reasons, even though it does have limitations. The perfect measurement assumption, for example, is entirely unreasonable even though it is commonly invoked in discussions of matching (and many, if not most, other methodological pieces). We rely on the random-sample perspective because we feel it is the most natural framing of these methods for the typical sociologist, even though many of the classic applications and early methodological pieces on matching do not reference random-sample surveys (instead relying on convenience and choice-based sampling). Pinning down the exact consequences of the assumed sampling scheme is important, as shown in Imbens (2004), for developing estimates of the expected variability of matching estimates. We discuss these issues in more detail in the penultimate section of this article.

3. See Winship and Morgan (1999) and Sobel (1995) for presentations of the counterfactual model in sociology. In this article, we adopt the foundational assumptions of the literature on counterfactual causality, such as the stable unit treatment value assumption, which stipulates that the causal effect for each individual does not depend on the treatment status of any other individual in the population. When this nonindependence assumption is violated, complications beyond the scope of this article arise.

4. For extensions of matching to multivalued causal/treatment variables, see Angrist and Krueger (1999), Hirano and Imbens (2004), Imbens (2000), Lechner (2002a, 2002b), Lu et al. (2001), Rosenbaum (2002), and Imai and van Dyk (2004). As one will see from reading this literature, the added complexity presented by multivalued and continuous treatments can be considerable, to the extent that matching loses much of its transparency and is then no more intuitive than regression (and, because of its unfamiliarity, then appears vastly more complex than regression). For these reasons, for the foreseeable future, we expect that most applied researchers will use matching only for the estimation of binary causal effects. Since such effects are usually the primitives of all more encompassing multivalued treatment effects, this may not be as severe of a limitation as one might fear.

5. There is a wide variety of notation in the potential outcome literature, and we have adopted notation that we feel is the easiest to grasp. Equation (1) is often written as one of the following alternatives: $\Delta_i = Y_{1i} - Y_{0i}$, $\delta_i = Y_i^t - Y_i^c$, $\delta_i = Y_{ti} - Y_{ci}$, $\tau_i = Y_i(1) - Y_i(0)$, or

variants thereof. We therefore use the right superscript to denote the potential treatment state of the corresponding potential outcome variable.

6. In other words, the subscript N serves the same basic notational function as an overbar on Y_i, as in $\bar{Y}_i$. We use this sub-N notation, as it allows for greater clarity in aligning sample and population-level conditional expectations for subsequent expressions.

7. Although this convergence notation may well be superfluous, we err on the side of precision of notation at this point because, as we show later, much of the confusion over the power of matching arises from a lack of appreciation for the different problems created by sparseness of data and sampling error relative to more serious forms of incomparability of treatment and control cases.

8. Although an ignorability assumption is satisfied in this case, it would be satisfied in weaker scenarios as well. As defined in Rubin (1978), ignorability of treatment assignment holds even if Y^1 and Y^0 are not fully independent of D but only independent of D after conditioning on observed variables that determine treatment selection. Rosenbaum and Rubin (1983a) then define strong ignorability to develop the matching literature. To Rubin's ignorability assumption, Rosenbaum and Rubin (1983a) required for strong ignorability that each subject have a nonzero probability of being assigned to either the treatment or the control group. Despite these clear definitions, the term *ignorability* is often defined in different ways in the literature. We suspect that this varied history of usage explains why Rosenbaum (2002) rarely uses the term in his monograph on observational data analysis, even though he is generally credited, along with Rubin, with developing the ignorability semantics in this literature. And it also explains why much of the most recent econometrics literature uses the words *unconfoundedness* and *exogeneity* for the same set of independence and conditional-independence assumptions (see Imbens 2004).

9. In the main text of this article, we generally refer to collections of variables with bold capital letters, such as **S**. For brevity, we rarely qualify these expressions as either vectors that exist for each individual (i.e., as k-by-1-column vectors of values on variables that exist for each individual) or as matrices that capture all values for all individuals (i.e., as n-by-k matrices with individuals as rows and variables as columns). Where such distinctions are important, we are more specific.

10. When in this situation, researchers often argue that the naive estimator is subject to bias (either generic omitted variable bias or individually generated selection bias). But since a perfect stratification of the data can be formulated, the study is said to be free of hidden bias (see Rosenbaum 2002), treatment assignment is ignorable, or treatment selection is on the observable variables **S** only (Heckman et al. 1999). Rosenbaum (2002) stresses the utility of asserting a no-hidden-bias assumption in an observational study but not then succumbing to overconfidence. The assumption allows one to obtain a causal effect estimate, but the initial estimate must be interpreted with caution and examined for its sensitivity to reasonable violations of Assumptions 1-S and 2-S.

11. This section shows one weakness of the random-sample survey setup that we have chosen as the background sampling framework for exposition. Since we rely on the convergence results stated earlier, it is now somewhat unnatural to just assert that the sample moments "equal" the population moments because the sample is sufficiently large. For purists, read *equal* in this section as equal in the asymptotic sense.

12. Note further that it is telling that we cannot think of a realistic sociological example that is as simple as this hypothetical example.

13. The naive estimator can be calculated for this example, and it would equal 8.05 for a very large sample because $[8(.325) + 14(.675)] - [2(.667) + 6(.167) + 10(.167)]$ is equal to 8.05. See the last row of Table 3 for the population analogs to the two pieces of the naive estimator.

14. As Rosenbaum (1987) later clarified (see also Rubin and Thomas 1996), the estimated propensity scores do a better job of balancing the observed variables in $\mathbf{S}$ than the true propensity scores would in any actual application since the estimated propensity scores correct for the chance imbalances in $\mathbf{S}$ that characterize any finite sample. This insight has led to a growing literature that seeks to balance variables in $\mathbf{S}$ by various computationally intensive but powerful nonparametric techniques. We discuss this literature later, and for now, we present only parametric models, as they dominate the foundational literature on matching.

15. The parameterization of Figure 1 is a constrained tensor product spline regression for the index function of a logit. See Ruppert, Wand, and Carroll (2003) for examples of such parameterizations. Figure 1 is generated by setting $\mathbf{S}_i\phi$ in equation (17) to $-2 + 3(A_i) - 3(A_i - .1) + 2(A_i - .3) - 2(A_i - .5) + 4(A_i - .7) - 4(A_i - .9) + 1(B_i) - 1(B_i - .1) + 2(B_i - .7) - 2(B_i - .9) + 3(A_i - .5)(B_i - .5) - 3(A_i - .7)(B_i - .7)$.

16. In effect, this setup establishes A and B as two independent multinomial distributions with equal probability mass for each of their respective 100 values.

17. We should note that one could easily generate an example where matching vastly outperforms linear regression by allowing Y_i^1 and Y_i^0 to be nonlinear in A_i and B_i.

18. All three regression estimators yield estimates that are typically interpreted as estimates of the average treatment effect, and thus we have placed them in the first column of the table (even though they could be regarded as estimators of other parameters as well). Notice that, as estimates of the treatment effect for the treated, they are on average too small.

19. The Mahalanobis metric is $(\mathbf{S}_i - \mathbf{S}_j)' \Sigma^{-1} (\mathbf{S}_i - \mathbf{S}_j)$, where Σ is the covariance matrix of the variables in $\mathbf{S}$ (usually calculated for the treatment cases only). There is a long tradition in this literature of using Mahalanobis matching in combination with propensity score matching. As Diamond and Sekhon (2005) note, propensity score matching balances the expectations of the variables in $\mathbf{S}$, and thereafter, Mahalanobis matching can be used to further balance the higher moments of the joint distribution of $\mathbf{S}$. This proposal is similar to what Rosenbaum (2002) advocates in some situations, and Diamond and Sekhon offer a genetic algorithm for pursuing this possibility, which we discuss later.

20. To estimate the treatment effect for the treated, the ranges of the variables in $\mathbf{S}$ must be the same for the treatment and control cases. We do not mention this requirement in the text, as there is a literature (see Heckman, Ichimura, and Todd 1997, 1998), which we discuss later, that defines the treatment effect for the treated on the common support and argues that this is often the central goal of analysis. Thus, even if the support of $\mathbf{S}$ is not the same in the treatment and control groups, an average treatment effect among a subset of the treated can be estimated.

21. There is an ignorability variant of this expectation-based assumption: Treatment assignment is independent of Y^0 conditional on $\mathbf{S}$.

22. One weakness of the traditional algorithm when used without replacement is that the estimate will vary depending on the initial ordering of the treatment cases. A second weakness is that without replacement, the sum distance for all treatment cases will generally not be the minimum because control cases that might make better matches to later treatment cases may be used early in the algorithm. See our discussion of optimal matching later.

23. A related form of matching, known as radius matching (see Dehejia and Wahba 2002), matches all control cases within a particular distance—the "radius"—from the treatment case and gives the selected control cases equal weight. If there are no control cases within the radius of a particular treatment case, then the nearest available control case is used as the match.

24. Increasing the bandwidth increases bias but lowers variance. Smith and Todd (2005) find that estimates are fairly insensitive to the size of the bandwidth.

25. Another criterion for choosing among alternative matching estimators is relative efficiency. Our reading of the literature suggests that little is known about the relative efficiency of these estimators (see especially Abadie and Imbens 2006; Hahn 1998; Imbens 2004), even though there are claims in the literature that kernel-based methods are the most efficient. The efficiency advantage of kernel-matching methods is only a clear guide to practice if kernel-based methods are known to be no more biased than alternatives. But the relative bias of kernel-based methods is application dependent and should interact further with the bandwidth of the kernel. Thus, it seems that we will only know for sure which estimators are most efficient for which types of applications when statisticians discover how to calculate the sampling variances of all alternative estimators. Thereafter, it should be possible to compute mean squared error comparisons across alternative estimators for sets of typical applications.

26. One method for matching on both the Mahalanobis metric and the propensity score is to include the propensity score in the Mahalanobis metric. A second is to use interval matching and divide the data into blocks using one metric and then match on the second metric within blocks.

27. To be precise, we generated a sample using a multinomial distribution from a race-by-region-by-urbanicity grid from the data in Morgan (2001). We then simulated socioeconomic status as random draws from normal distributions, with means and standard deviations estimated separately for each of the race-by-region-by-urbanicity cells using the data from Morgan. Then, we generated all other variables iteratively, building on top of these variables, using joint distributions (where possible) based on estimates from the National Education Longitudinal Study (NELS) data. Since we relied on standard parametric distributions, the data are somewhat more smooth than the original NELS data (which thereby gives an advantage to parametric regression relative to nonparametric matching methods, as we note later).

28. The index of the assumed logit was $-4.6 - .69$(Asian) $+ .23$(Hispanic) $- .76$(black) $- .46$(Native American) $+ 2.7$(urban) $+ 1.5$(northeast) $+ 1.3$(north central) $+ .35$(south) $- .02$(siblings) $- .018$(bedroom) $+ .31$(two parents) $+ .39$(socioeconomic status) $+ .33$(cognitive skills) $- .032$(socioeconomic status squared) $- .23$(cognitive skills squared) $- .084$(socioeconomic status)(cognitive skills) $- .37$(two parents)(black) $+ 1.6$(northeast)(black) $- .38$(north central)(black) $+ .72$(south)(black) $+ .23$(two parents)(Hispanic) $- .74$(northeast)(Hispanic) $- 1.3$(north central)(Hispanic) $- 1.3$(south)(Hispanic) $+ .25$(individual treatment effect $-$ average treatment effect).

29. We do not provide a review of software routines, as such a review would be immediately out of date upon publication. At present, three additional sets of routines seem to be in use in the applied literature (see Hansen 2004b; Ho et al. 2004; Sekhon 2005).

30. It is noteworthy that even when we requested equivalent matching estimates from alternative software routines (even beyond those presented in Table 6), we obtained different

estimates. We cannot determine the source of these differences from the documentation provided by the software's creators.

31. At the same time, this sort of example shows that even our earlier definition of a "perfect stratification" is somewhat underspecified. According to the definition stated earlier, if self-selection on the causal effect occurs, a perfect stratification is available only if variables that accurately measure anticipation of the causal effect for each individual are also available and duly included in **S**. Thus, perhaps it would be preferable to refer to three types of perfect stratification: one where Assumption 1-S is valid (which enables estimation of the average treatment effect for the untreated), one where Assumption 2-S is valid (which enables estimation of the average treatment for the treated), and one where both are valid (which enables estimation of the average treatment effect, as well as the average treatment effects for the treated and the untreated).

32. One could also estimate these three average treatment effects for hypothetical Example 1 in the same way, although it would require specifying *S* as two dummy variables (rather than one interval-scaled variable) when estimating the propensity score. Thus, even though we stressed the weighting of stratified estimates in that example, we could also have showed how individual-level weighting via estimated propensity scores is also consistent for each of the three treatment effects.

33. Rubin (1977) provides simple and elegant examples of all such complications, highlighting the importance of assumptions about the relationships between covariates and outcomes (see also Holland and Rubin 1983; Rosenbaum 1984).

34. *Support* is often given slightly different definitions depending on the context, although most definitions are consistent with a statement such as the following: the union of all infinitesimally small intervals of a probability distribution that have true nonzero probability mass.

35. As argued by Heckman and Vytlacil (1999, 2000, 2004), these types of treatment effect estimates are among the most informative, both for policy guidance and theoretical prediction, as they focus on those at the margin of treatment participation (or causal exposure).

36. Two of the three matching software routines that we used for Example 4 allow one to calculate bootstrapped standard errors in STATA. This is presumably because these easy-to-implement methods were once thought to provide a general framework for estimating the standard errors of alternative matching estimators and hence were a fair way to compare the relative efficiency of alternative matching estimators (see Tu and Zhou 2002). Unfortunately, Abadie and Imbens (2004) show that conventional bootstrapping is fragile and will not work in general for matching estimators. Whether generalized forms of bootstrapping may still be used effectively remains to be determined.

37. There is also a related set of randomization inference techniques, built up from consideration of all of the possible permutations of treatment assignment patterns that could theoretically emerge from alternative enactments of the same treatment assignment routine (see Rosenbaum 2002). These permutation ideas generate formulas for evaluating specific null hypotheses, which, from our perspective, are largely uncontroversial. They are especially reasonable when the analyst has deep knowledge of a relatively simple treatment assignment regime and has reason to believe that treatment effects are constant in the population. Although Rosenbaum (2002) provides large-sample approximations for these permutation-based tests, the connections to the recent econometrics literature that draws on nonparametric convergence results have not yet been established.

References

Abadie, Alberto. 2002. "Bootstrap Tests for Distributional Treatment Effect in Instrumental Variable Models." *Journal of the American Statistical Association* 97:284-92.

Abadie, Alberto, David Drukker, Jane L. Herr, and Guido W. Imbens. 2001. "Implementing Matching Estimators for Average Treatment Effects in Stata." *The Stata Journal* 1:1-18.

Abadie, Alberto and Guido W. Imbens. 2004. "On the Failure of the Bootstrap for Matching Estimators." Working paper, John F. Kennedy School of Government, Harvard University.

———. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica*. 74:235-67.

Althauser, Robert P. and Donald B. Rubin. 1970. "The Computerized Construction of a Matched Sample." *American Journal of Sociology* 76:325-46.

———. 1971. "Measurement Error and Regression to the Mean in Matched Samples." *Social Forces* 50:206-14.

Angrist, Joshua D. and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." Pp. 1277-1366 in *Handbook of Labor Economics*, vol. 3, edited by O. C. Ashenfelter and D. Card. Amsterdam: Elsevier.

Becker, Sascha O. and Andrea Ichino. 2002. "Estimation of Average Treatment Effects Based on Propensity Scores." *The Stata Journal* 2:358-77.

Berk, Richard A. and Phyllis J. Newton. 1985. "Does Arrest Really Deter Wife Battery? An Effort to Replicate the Findings of the Minneapolis Spouse Abuse Experiment." *American Sociological Review* 50:253-62.

Berk, Richard A., Phyllis J. Newton, and Sarah Fenstermaker Berk. 1986. "What a Difference a Day Makes: An Empirical Study of the Impact of Shelters for Battered Women." *Journal of Marriage and Family* 48:481-90.

Cochran, William G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24:295-313.

Dehejia, Rajeev H. and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94:1053-62.

———. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84:151-61.

Diamond, Alexis and Jasjeet S. Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." Working paper, Travers Department of Political Science, UC Berkeley.

DiPrete, Thomas A. and Henriette Engelhardt. 2004. "Estimating Causal Effects With Matching Methods in the Presence and Absence of Bias Cancellation." *Sociological Methods & Research* 32:501-28.

DiPrete, Thomas A. and Markus Gangl. 2004. "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation With Imperfect Instruments." *Sociological Methodology* 34:271-310.

Freedman, Ronald and Amos H. Hawley. 1949. "Unemployment and Migration in the Depression." *Journal of the American Statistical Association* 44:260-72.

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

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66:315-31.

Ham, J. C., X. Li, and P. B. Reagan. 2003. "Propensity Score Matching, a Distance-Based Measure of Migration, and the Wage Growth of Young Men." Working paper, Department of Sociology and Center for Human Resource Research, Ohio State University.

Hansen, Ben B. 2004a. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association* 99:609-18.

———. 2004b. "Optmatch, an Add-on Package for R." Department of Statistics, University of Michigan.

Harding, David J. 2003. "Counterfactual Models of Neighborhood Effects: The Effect of Neighborhood Poverty on Dropping Out and Teenage Pregnancy." *American Journal of Sociology* 109:676-719.

Heckman, James J. 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." *Quarterly Journal of Economics* 115:45-97.

Heckman, James J., Hidehiko Ichimura, Jeffery A. Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66:1017-98.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence From Evaluating a Job Training Programme." *Review of Economic Studies* 64:605-54.

———. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65:261-94.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." Pp. 1865-2097 in *Handbook of Labor Economics*, vol. 3, edited by O. C. Ashenfelter and D. Card. Amsterdam: Elsevier.

Heckman, James J. and Edward Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences of the United States of America* 96:4730-34.

———. 2000. "The Relationship Between Treatment Parameters Within a Latent Variable Framework." *Economics Letters* 66:33-9.

———. 2004. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73:669-738.

Hirano, Keisuke and Guido W. Imbens. 2004. "The Propensity Score With Continuous Treatments." Pp. 73-84 in *Applied Bayesian Modeling and Causal Inference From Incomplete-Data Perspectives: An Essential Journey With Donald Rubin's Statistical Family*, edited by A. Gelman and X.-L. Meng. New York: John Wiley.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71:1161-89.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2004. "Matchit." http://gking.harvard.edu/matchit

Hoffer, Thomas, Andrew M. Greeley, and James S. Coleman. 1985. "Achievement Growth in Public and Catholic Schools." *Sociology of Education* 58:74-97.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945-70.

Holland, Paul W. and Donald B. Rubin. 1983. "On Lord's Paradox." Pp. 3-25 in *Principles of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*, edited by H. Wainer and S. Messick. Hillsdale, NJ: Lawrence Erlbaum.

Imai, Kosuke and David A. van Dyk. 2004. "Causal Inference With General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99:854-66.

Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87:706-10.

———. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86:4-29.

Lechner, Michael. 2002a. "Some Practical Issues in the Evaluation of Heterogeneous Labour Market Programmes by Matching Methods." *Journal of the Royal Statistical Society* 165:59-82.

———. 2002b. "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies." *Review of Economics and Statistics* 84:205-20.

Leuven, Edwin and Barbara Sianesi. 2003. "Psmatch2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing.'' http://ideas.repec.org/c/boc/bocode/s432001.html, version x.x.x.

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

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Morgan, Stephen L. 2001. "Counterfactuals, Causal Effect Heterogeneity, and the Catholic School Effect on Learning." *Sociology of Education* 74:341-74.

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

Rosenbaum, Paul R. 1984. "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society, Series A* 147:656-66.

———. 1987. "Model-Based Direct Adjustment." *Journal of the American Statistical Association* 82:387-94.

———. 1989. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association* 84:1024-32.

———. 1991. "Sensitivity Analysis for Matched Case Control Studies." *Biometrics* 47:87-100.

———. 1992. "Detecting Bias With Confidence in Observational Studies." *Biometrika* 79:367-74.

———. 2002. *Observational Studies*. New York: Springer.

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

———. 1983b. "Assessing Sensitivity to an Unobserved Covariate in an Observational Study With Binary Outcome." *Journal of the Royal Statistical Society* 45:212-8.

———. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79:516-24.

———. 1985a. "Constructing a Control Group Using Multivariate Matched Sampling Methods." *The American Statistician* 39:33-8.

———. 1985b. "The Bias Due to Incomplete Matching." *Biometrics* 41:103-16.

Rubin, Donald B. 1973a. "Matching to Remove Bias in Observational Studies." *Biometrics* 29:159-83.

———. 1973b. "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies." *Biometrics* 29:185-203.

———. 1976a. "Multivariate Matching Methods That Are Equal Percent Bias Reducing, I: Some Examples." *Biometrics* 32:109-20.

———. 1976b. "Multivariate Matching Methods That Are Equal Percent Bias Reducing, II: Maximums on Bias Reduction for Fixed Sample Sizes." *Biometrics* 32:121-32.
———. 1977. "Assignment to Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2:1-26.
———. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6:34-58.
———. 1979. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74:318-28.
———. 1980. "Bias Reduction Using Mahalanobis-Metric Matching." *Biometrics* 36:293-8.
Rubin, Donald B. and Neal Thomas. 1996. "Matching Using Estimated Propensity Scores: Relating Theory to Practice." *Biometrics* 52:249-64.
———. 2000. "Combining Propensity Score Matching With Additional Adjustments for Prognostic Covariates." *Journal of the American Statistical Association* 95:573-85.
Ruppert, David, M. P. Wand, and Raymond J. Carroll. 2003. *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.
Sekhon, Jasjeet. 2005. "Matching." http://sekhon.polisci.berkeley.edu/
Smith, Herbert L. 1997. "Matching With Multiple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology* 27:325-53.
Smith, Jeffery A. and Petra Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125:305-53.
Sobel, Michael E. 1995. "Causal Inference in the Social and Behavioral Sciences." Pp. 1-38 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by G. Arminger, C. C. Clogg, and M. E. Sobel. New York: Plenum.
Tu, Wanzhu and Xiao-Hua Zhou. 2002. "A Bootstrap Confidence Interval Procedure for the Treatment Effect Using Propensity Score Subclassification." *Health Services & Outcomes Research Methodology* 3:135-47.
van der Laan, Mark J. and James M. Robins. 2003. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.
Winship, Christopher and Stephen L. Morgan. 1999. "The Estimation of Causal Effects From Observational Data." *Annual Review of Sociology* 25:659-706.
Yinger, Milton J., Kiyoshi Ikeda, and Frank Laycock. 1967. "Treating Matching as a Variable in a Sociological Experiment." *American Sociological Review* 32:801-12.

Stephen L. Morgan, PhD, is an associate professor of sociology and the director of the Center for the Study of Inequality at Cornell University. His areas of interest include social stratification, sociology of education, and quantitative methodology. He is the author of the 2005 book *On the Edge of Commitment: Educational Attainment and Race in the United States*, published by Stanford University Press.

David J. Harding, PhD, is an NICHD postdoctoral fellow at the Population Studies Center at the University of Michigan. Beginning fall 2006, he will be an assistant professor in the Department of Sociology and assistant research scientist in the Population Studies Center at the University of Michigan. His current interests include urban poverty, education, adolescent romantic and sexual behavior, and methodology.