CHAPTER 1

Introduction

Propensity score analysis is a class of statistical methods developed for estimating treatment effects with nonexperimental or observational data. Specifically, propensity score analysis offers an approach to program evaluation when randomized trials are infeasible or unethical, or when researchers need to assess treatment effects from survey data, census data, administrative data, medical records data, or other types of data "collected through the observation of systems as they operate in normal practice without any interventions implemented by randomized assignment rules" (Rubin, 1997, p. 757). In the social and health sciences, researchers often face a fundamental task of drawing conditioned casual inferences from quasi-experimental studies. Analytical challenges in making causal inferences can be addressed by a variety of statistical methods, including a range of new approaches emerging in the field of propensity score analysis.

This book focuses on seven closely related but technically distinct models for estimating treatment effects: (1) Heckman's sample selection model (Heckman, 1976, 1978, 1979) and its revised version (Maddala, 1983), (2) propensity score matching (Rosenbaum, 2002b; Rosenbaum & Rubin, 1983) and related models, (3) propensity score subclassification (Rosenbaum & Rubin, 1983, 1984), (4) propensity score weighting (Hirano & Imbens, 2001; Hirano, Imbens, & Ridder, 2003; McCaffrey, Ridgeway, & Morral, 2004), (5) matching estimators (Abadie & Imbens, 2002, 2006), (6) propensity score analysis with nonparametric regression (Heckman, Ichimura, & Todd, 1997, 1998), and (7) propensity score analysis of categorical or continuous treatments (Hirano & Imbens, 2004; Imbens, 2000; Joffe & Rosenbaum, 1999).

Although statisticians and econometricians have not reached consensus on the scope and content of propensity score analysis, the statistical models described in this book share several similar characteristics: Each has the objective of assessing treatment effects and controlling for covariates, each represents state-of-the-art analysis in program evaluation, and each can be employed to overcome various kinds of challenges encountered in research.

Although the randomized controlled trial is deemed to be the gold standard in research design, true experimental designs are not always possible, practical, or even desirable in the social and health sciences. Given a continuing reliance on quasi-experimental design, researchers have increasingly sought methods to improve estimates of program effects.

Over the past 35 years, methods of program evaluation have undergone a significant change as researchers have recognized the need to develop more efficient approaches for assessing treatment effects from studies based on observational data and for evaluations based on quasi-experimental designs. This growing interest in seeking consistent and efficient estimators of program effectiveness led to a surge in work focused on estimating average treatment effects under various sets of assumptions. Statisticians (e.g., Rosenbaum & Rubin, 1983) and econometricians (e.g., Heckman, 1978, 1979) have made substantial contributions by developing and refining new approaches for the estimation of causal effects from observational data. Collectively, these approaches are known as *propensity score analysis*.

Econometricians have integrated propensity score models into other econometric models (i.e., instrumental variable, control function, difference-in-differences estimators) to perform less expensive and less intrusive nonexperimental evaluations of social, educational, and health programs. Furthermore, recent criticism and reformulations of the classical experimental approach in econometrics symbolize an important shift in evaluation methods. The significance of this movement was evidenced by the selection of James Heckman as one of the 2000 Nobel Prize award winners in the field of economics. The prize recognized his development of theory and methods for data analysis in selective samples.

As a new and rapidly growing class of evaluation methods, propensity score analysis is by no means conceived as the best alternative to randomized experiments. In empirical research, it is still unknown under what circumstances the approach appears to reduce selection bias and under what circumstances the conventional regression approach (i.e., use of statistical controls) remains adequate. There are certainly debates about the advantages and disadvantages of propensity score modeling. These focus, primarily, on the extent to which propensity score methods offer effective and efficient estimates of treatment effects and on the degree to which they help address many challenging issues embedded in program evaluation, policy evaluation, and causal inference. The call for developing and using strong research designs to provide a comprehensive understanding of causal processes in program evaluation remains a paramount challenge in all fields of practice. However, it is also a consensus among prominent researchers that the propensity score approach has reached a *mature* level. For instance, Imbens and Wooldridge (2009) evaluated recent developments in the econometrics of program evaluation, primarily the methods described by this book, and concluded that

> at this stage, the literature has matured to the extent that it has much to offer the empirical researchers. Although the evaluation problem is one where identification problems are important, there is currently a much better understanding of which assumptions are most useful, as well as a better set of methods for inference given different sets of assumptions. (p. 8)

Representing the interest in—and indeed perceived utility of—these new methods, the propensity score approach has been employed in a variety of disciplines and professions such as education (Morgan, 2001), epidemiology (Normand et al., 2001), medicine (e.g., Earle et al., 2001; Gum, Thamilarasan, Watanabe, Blackstone, & Lauer, 2001), psychology (Jones, D'Agostino, Gondolf, & Heckert, 2004), social work (Barth, Greeson, Guo, & Green, 2007;

Barth, Lee, Wildfire, & Guo, 2006; Guo, Barth, & Gibbons, 2006; Weigensberg, Barth, & Guo, 2009), and sociology (H. L. Smith, 1997). In social welfare studies, economists and others used propensity score methods in evaluations of the National Job Training Partnership Act program (Heckman, Ichimura, & Todd, 1997), the National Supported Work Demonstration (LaLonde, 1986), and the National Evaluation of Welfare-to-Work Strategies Study (Michalopoulos, Bloom, & Hill, 2004).

In describing these new methods, the preparation and writing of this book was guided by two primary objectives. The first objective was to introduce readers to the origins, main features, and debates centering on the seven models of propensity score analysis. We hope this introduction will help accomplish our second objective of illuminating new ideas, concepts, and approaches that social and health sciences researchers can apply to their own fields to solve problems they might encounter in their research efforts. In addition, this book has two overarching goals. Our primary goal is to make the past three decades of theoretical and technological advances in analytic methods accessible and available in a less technical and more practical fashion. The second goal is to promote discussions among social and health sciences researchers regarding emerging strategies and methods for estimating causal effects using nonexperimental methods.

The aim of this chapter is to provide an overview of the propensity score approach. Section 1.1 presents a definition of observational study. Section 1.2 reviews the history and development of the methods. Section 1.3 is an overview of the randomized experimental approach, which is the gold standard developed by statisticians and the model that should serve as a foundation for the nonexperimental approach. Section 1.4 offers examples drawn from literature beyond the fields of econometrics and statistics. These examples are intended to help readers determine the situations in which the propensity score approach may be appropriate. Section 1.5 reviews the computing software packages that are currently available for propensity score analysis and the main features of the package used in the models presented throughout this book. Section 1.6 outlines the organization of the book.

1.1 OBSERVATIONAL STUDIES

The statistical methods we discuss may be generally categorized as methods for *observational studies.* According to Cochran (1965), an observational study is an empirical investigation whose objective is to elucidate causal relationships (i.e., cause and effect) when it is infeasible to use controlled experimentation and to assign participants at random to different procedures.

In the general literature related to program evaluation (i.e., nonstatistically oriented literature), researchers use the term *quasi-experimental* more frequently than *observational* studies, with the term defined as studies that compare groups but lack the critical element of random assignment. Indeed, *quasi-experiments* can be used interchangeably with *observational studies,* as described in the following quote from Shadish, Cook, and Campbell (2002):

> Quasi-experiments share with all other experiments a similar purpose—to test descriptive causal hypotheses about manipulable causes—as well as many

> structural details, such as the frequent presence of control groups and pretest measures, to support a counterfactual inference about what would have happened in the absence of treatment. But, by definition, quasi-experiments lack random assignment. Assignment to conditions is by means of *self-selection,* by which units choose treatment for themselves, or means of *administrator selection,* by which teachers, bureaucrats, legislators, therapists, physicians, or others decide which persons should get which treatment. (pp. 13–14)

Two features of observational studies merit particular emphasis. First, an observational study concerns treatment effects. A study without a treatment—often called an *intervention* or a *program*—is neither an experiment nor an observational study. Most public opinion polls, forecasting efforts, investigations of fairness and discrimination, and many other important empirical studies are neither experiments nor observational studies (Rosenbaum, 2002b). Second, observational studies can employ data from nonexperimental, nonobservational studies as long as the focus is on assessing treatment or the effects of receiving a particular service. By this definition, observational data refer to data that were generated by something other than a randomized experiment and typically include surveys, censuses, or administrative records (Winship & Morgan, 1999).

1.2 HISTORY AND DEVELOPMENT

The term *propensity score* first appeared in a 1983 article by Rosenbaum and Rubin, who described the estimation of causal effects from observational data. Heckman's (1978, 1979) work on dummy endogenous variables using simultaneous equation modeling addressed the same issue of estimating treatment effects when assignment was nonrandom; however, Heckman approached this issue from a perspective of sample selection. Although Heckman's work on the dummy endogenous variable problem employed different terminology, he used the same approach toward estimating a participant's probability of receiving one of two conditions. Both schools of thought (i.e., the econometric tradition of Heckman and the statistical tradition of Rosenbaum and Rubin) have had a significant influence on the direction of the field, although the term *propensity score analysis,* coined by Rosenbaum and Rubin, is used more frequently as a general term for the set of related techniques designed to correct for selection bias in observational studies.

The development of the propensity score approach signified a convergence of two traditions in studying causal inferences: the econometric tradition that primarily relies on structural equation modeling and the statistical tradition that primarily relies on randomized experiments (Angrist, Imbens, & Rubin, 1996; Heckman, 2005). The econometric tradition dates back to Trygve Haavelmo (1943, 1944), whose pioneering work developed a system of linear simultaneous equations that allowed analysts to capture interdependence among outcomes, to distinguish between fixing and conditioning on inputs, and to parse out true causal effects and spurious causal effects. The task of estimating *counterfactuals*, a term generally developed and used by statisticians, is explored by econometricians in the form of a *switching regression model* (Maddala, 1983; Quandt, 1958, 1972). Heckman's (1978, 1979) development of a two-step estimator is credited as the field's pioneering work in explicitly modeling the causes of selection in the form of a dummy endogenous variable. As previously mentioned, Heckman's work followed econometric conventions and solved the problem through structural equation modeling.

Historically quite distinct from the econometric tradition, the statistical tradition can be traced back to Fisher (1935/1971), Neyman (1923), and Rubin (1974, 1978). Unlike conventions based on linear simultaneous equations or structural equation models, the statistical tradition is fundamentally based on the randomized experiment. The principal notion in this formulation is the study of *potential outcomes,* known as the *Neyman-Rubin counterfactual framework*. Under this framework, the causal effects of treatment on sample participants (already exposed to treatments) are explored by observing outcomes of participants in samples not exposed to the treatments. Rubin extended the counterfactual framework to more complicated situations, such as observational studies without randomization.

For a detailed discussion of these two traditions, readers are referred to a special issue of the *Journal of the American Statistical Association* (1996, Vol. 91, No. 434), which presents an interesting dialogue between statisticians and econometricians. Significant scholars in the field—including Greenland, Heckman, Moffitt, Robins, and Rosenbaum—participated in a discussion of a study that used instrumental variables to identify causal effects, particularly the local average treatment effect (Angrist et al., 1996).

It is worth noting that the development of propensity score models did not occur in isolation from other important developments. At the same time that propensity score methods were emerging, the social, behavioral, and health sciences witnessed progress in the development of other statistical methods, such as methods for the control of clustering in multilevel data—for instance, the linear mixed model (Laird & Ware, 1982), hierarchical linear modeling (Raudenbush & Bryk, 2002), and robust standard error estimator (Huber, 1967; White, 1980); methods to analyze latent variables and to model complex structural relationships among latent variables (e.g., analyzing moderating as well as mediating effects, or models to depict nonrecursive relationship between latent variables)—that is, the structural equation modeling (Bollen, 1989; Jöreskog, 1971); methods for analyzing categorical and limited dependent variables—that is, the generalized linear models (Nelder & Wedderburn, 1972); methods for analyzing time-to-event data—for instance, the proportional hazards model (Cox, 1972) and marginal approaches to clustered event data (Lee, Wei, & Amato, 1992; Wei, Lin, & Weissfeld, 1989); and more. When researchers are engaged in observational studies, many of these newly developed models need to be applied in conjunction with propensity score methods, and by the same token, a successful propensity score analysis always requires a careful examination of other issues of data analysis, including addressing potential violations of statistical assumptions by employing these newly developed methods. In this book, whenever possible, we describe the application of propensity score models in settings where other data issues are present, and we show how to employ propensity score models in conjunction with the application of other statistical approaches.

1.3 RANDOMIZED EXPERIMENTS

The statistical premise of program evaluation is grounded in the tradition of the randomized experiment. Therefore, a natural starting point in a discussion of causal attribution in observational studies is to review key features of the randomized experiment. According to Rosenbaum (2002b), a theory of observational studies must have a clear conceptual linkage to randomization, so that the consequences of the absence of randomization can be understood. For example, sensitivity analysis is among Rosenbaum's approaches to handling data with hidden selection bias; this approach includes the use of test statistics that were developed primarily for randomized experiments, such as Wilcoxon's signed rank statistic and Hodges-Lehmann estimates. However, the critiques of social experiments by econometricians (e.g., Heckman & Smith, 1995) frequently include description of the conditions under which randomization is infeasible, particularly under the setting of social behavioral research. Thus, it is important to review principles and types of randomized experiments, randomization tests, and the challenges to this tradition. Each of these topics is addressed in the following sections.

1.3.1 Fisher's Randomized Experiment

The invention of the randomized experiment is generally credited to Sir Ronald Fisher, one of the foremost statisticians of the 20th century. Fisher's book, *The Design of Experiments* (1935/1971), introduced the principles of randomization, demonstrating them with the now-famous example of testing a British woman's tea-tasting ability. This example has been cited repeatedly to illustrate the power of randomization and the logic of hypothesis testing (see, e.g., Maxwell & Delaney, 1990; Rosenbaum, 2002b). In a somewhat less technical fashion, we include this example as an illustration of important concepts in randomized experimentation.

In Fisher's (1935/1971) words, the problem is as follows:

> A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested. (p. 11)

During Fisher's time, the dominant practice in experimentation was to control covariates or confounding factors that might contaminate treatment effects. Therefore, to test a person's tasting ability (i.e., the true ability to discriminate two methods of tea preparation), a researcher would control factors that could influence the results, such as the temperature of tea, the strength of the tea, the use of sugar, and the amount of milk added, in addition to the myriad potential differences that might occur among the cups of tea used in an experiment. As Maxwell and Delaney (1990) pointed out,

> The logic of experimentation up until the time of Fisher dictated that to have a valid experiment here all the cups to be used "must be exactly alike," except for the independent variable being manipulated. Fisher rejected this dictum on two grounds. First, he argued that it was logically impossible to achieve, both in the

> example and in experimentation in general. . . . Second, Fisher argued that, even if it were conceivable to achieve "exact likeness," or more realistically, "imperceptible difference" on various dimensions of the stimuli, it would in practice be too expensive to attempt. (p. 40)

Instead of controlling for every potential confounding factor, Fisher proposed to *control for nothing,* namely, to employ a method of randomization. Fisher (1935/1971) described his design as follows:

> Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of what the test will consist, namely that she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such a manipulation. Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received. (p. 12)

Before going further, it is crucial to note several important points regarding Fisher's design. First, in this example, the unit of analysis is not individual ($N \neq 1$), but rather the presentation of the tea cups to the tea taster (i.e., $N = 8$, in which a total of 8 cases comprise the sample). Second, there is a treatment assignment process in this example, namely, the order of presentation of tea cups. Using Rosenbaum's (2002b) notation, this is a random variable **Z,** and any specific presentation of the tea cups to the taster is a realization of **Z,** or $\mathbf{Z} = \mathbf{z}$. For instance, if a specific presentation of the tea cups consists of four cups with milk added first followed by four cups with tea added first, then we may write $\mathbf{z} = (11110000)$, where **z** is just one of many possible assignments. In Rosenbaum's notation, these possible treatment assignments form a set of Ω, and $\mathbf{z} \in \Omega$. Determining the total number of elements in Ω (which Rosenbaum denoted as K) is an important task for experimental design and a task that can be accomplished using probability theory. This point will be discussed in more detail elsewhere. Third, there is an actual outcome **r,** which is the result of tasting the eight cups of tea. If the taster gives exactly the same order of tea cups as in the treatment assignment (i.e., she correctly identifies the first four cups as having the milk added first and the next four cups as having the tea added first), then the outcome would be recorded as $\mathbf{r} = (11110000)$. Last, the test essentially aims to determine whether the tea taster had the true ability to discriminate the two kinds of tea or whether she made her correct judgment accidentally by guessing. Thus, the null hypothesis (H_0) under testing would be "She has no ability to discriminate," and the test involves finding statistical evidence to reject the null hypothesis at a given significance level.

Building on these explanations, we continue with the tea-tasting test and describe how Fisher implemented his randomized experiment. One important feature of randomized experiments is that, in advance of implementation, the researcher must calculate probable outcomes for each study unit. Fisher (1935/1971) emphasized "forecasting all possible outcomes," even at a design stage when outcome data are completely absent: "In considering the appropriateness of any proposed experimental design, it is always needful to forecast all possible results of the experiment, and to have decided without ambiguity what interpretation shall be placed upon each one of them" (p. 12).

The key of such calculation is to know the total number of elements in the set of Ω (i.e., the value of K). In the above example, we simply made an arbitrary example of treatment assignment 11110000, although many other treatment assignments can be easily figured out, such as alternating cups of tea with the milk added first with the cups prepared by adding the tea infusion first (i.e., 10101010), or presenting four cups with tea infusion added first and then four cups with milk added first (i.e., 00001111). In statistics, the counting rules (i.e., permutations and combinations) inform us that the number of total possible ways to present the eight cups can be solved by finding out the number of combinations of eight things taken four at a time, or ${}_8C_4$, as

$${}_nC_r = \frac{n(n-1)(n-2)\ldots(n-r+1)}{r(r-1)(r-2)\ldots 1} = \frac{n!}{r!(n-r)!}.$$

The solution to our problem is

$$k = {}_8C_4 = \frac{8!}{4!4!} = 70.\text{[1]}$$

Therefore, there are 70 possible ways to present the tea taster with four cups with milk added first and four cups with tea added first. We can keep writing 11110000, 10101010, 00001111, . . . until we exhaust all 70 ways. Here, 70 is the number of total elements in the set of Ω or all possibilities for a treatment assignment.

To perform a statistical test of "H_0: No ability," Fisher turned to the task of looking into the possible outcomes **r**. Furthermore, if we define the taster's true ability to taste discriminately as requiring all eight cups she identified to match *exactly* what we presented to her, we can then calculate the probability of having the true outcome. The significance test performed here involves rejecting the null hypothesis, and the null hypothesis is expressed as "no ability." Fisher used the logic of "guessing the outcome right"; that is, the taster has no ability to discriminate but makes her outcome correct by guessing. Thus, what is the probability of having the outcome **r** that is identical to the treatment assignment **z?** The outcome **r** should have one set of values from the 70 possible treatment assignments; that is, the taster could guess any outcome from 70 possible outcomes of 11110000, 10101010, 00001111, ... Therefore, the probability of guessing the right outcome is 1/70 = .0124, which is a very low probability. Now, we can reject the null hypothesis under a small probability of making a Type I error (i.e., the tea taster did have the ability, but we erroneously rejected the "no ability" hypothesis), and the chance is indeed very low (.0124). In other words, based on statistical evidence (i.e., an examination of all possible outcomes), we can reject the "no ability" hypothesis at a statistical significance level of .05. Thus, we may conclude that under such a design, the taster may have true tasting ability ($p < .05$).

Rosenbaum (2002b) used $t(\mathbf{Z}, \mathbf{r})$ to denote the test statistic. In the preceding test scenario, we required a perfect match—a total of eight agreements—between the treatment (i.e., the order of tea cups presented to the tea-tasting expert) and the outcome (i.e., the actual outcome identified by the taster); therefore, the problem is to find out the probability $\{t(\mathbf{Z}, \mathbf{r}) > 8\}$. This probability can be more formally expressed in Rosenbaum's notation as follows:

$$\text{prob}\{t(\mathbf{Z},\mathbf{r}) \geq T\} = \frac{|\{z \in \Omega : t(Z,r) \geq T\}|}{K}, \text{or prob}\{t(\mathbf{Z},\mathbf{r}) \geq 8\} = \frac{1}{70} = .0124.$$

However, if the definition of "true ability" is relaxed to allow for six exact agreements rather than eight agreements (i.e., six cups in the order of outcome match to the order of presentation), we can still calculate the probability or significance in testing the null hypothesis of "no ability." As in the earlier computation, this calculation involves the comparison of actual outcome **r** to the treatment assignment **z,** and the tea taster's outcome could be any one of 70 possible outcomes. Let us assume that the taster gives her outcome as **r** = (11110000). We now need to examine how many treatment assignments (i.e., number of **z**) match this outcome under the relaxed definition of "true ability." The answer to this question is one perfect match (i.e., the match with eight agreements) plus 16 matches with six agreements (Rosenbaum, 2002b, p. 30), for a total of 17 treatment assignments. To illustrate, we provide all 17 treatment assignments that match to the taster's outcome 11110000:

perfect match **11110000,** and the following assignments with six exact agreements:

0**111**1**000**, 0**111000**1, 0**11100**1**0**, 0**1110**1**00**, **1**0**110**1**00**, **1**0**1100**1**0**, **1**0**11000**1, **1**0**11**1**000**, **11**0**100**1**0**, **11**0**1000**1, **11**0**10**1**00**, **11**0**1**1**000**, **111**0**000**1, **111**0**00**1**0**, **111**0**0**1**00**, **111**01**000**,

where bold numbers indicate agreements.[2]

Thus, the probability of having six exact agreements is 17/70 = .243. In Rosenbaum's notation, the calculation is

$$\text{prob}\{t(\mathbf{Z},\mathbf{r}) \geq T\} = \frac{|\{z \in \Omega : t(\mathbf{Z},\mathbf{r}) \geq T\}|}{K}, \text{or prob}\{t(\mathbf{Z},\mathbf{r}) \geq 6\} = \frac{17}{70} = .243.$$

That is, if we define "true ability" as correctly identifying six out of eight cups of tea, the probability of having a correct outcome increases to .243. The null hypothesis cannot be rejected at a .05 level. In other words, under this relaxed definition, we should be more conservative, or ought to be more reluctant, to declare that the tea taster has true ability. With a sample of eight cups in total and a relaxed definition of "ability," the statistical evidence is simply insufficient for us to reject the null hypothesis, and, therefore, the experimental design is less significant in testing true tasting ability.

We have described Fisher's famous example of randomized experiment in great detail. Our purpose of doing so is twofold. The first is to illustrate the importance of understanding two processes in generating intervention data: (1) the treatment assignment process (i.e., there is a random variable **Z,** and the total number of possible ways K is inevitably large) makes it possible to know in advance the probability of receiving treatment in a uniform randomized experiment and (2) the process of generating outcome data (i.e., there is an outcome variable **r**). This topic is revisited both in Chapters 2 and 3, in the discussion of the so-called ignorable treatment assignment, and in Chapter 11, in the discussion of selection bias and sensitivity analysis. The second purpose in providing a detailed description of Fisher's experiment was to call attention to the core elements of randomized experiments. According to Rosenbaum (2002b),

> First, experiments do not require, indeed cannot reasonably require, that experimental units be homogeneous, without variability in their responses. . . . Second, experiments do not require, indeed, cannot reasonably require, that experimental units be a random sample from a population of units. . . . Third, for valid inference about the effects of a treatment on the units included in an experiment, it is sufficient to require that treatments be allocated at random to experimental units—these units may be both heterogeneous in their responses and not a sample from a population. Fourth, probability enters the experiment only through the random assignment of treatments, a process controlled by the experimenter. (p. 23)

1.3.2 Types of Randomized Experiments and Statistical Tests

Fisher's framework laid the foundation for randomized experimental design. The method has become a gold standard in program evaluation and continues to be an effective and robust means for assessing treatment effects in nearly every field of interest from agriculture and business, to computer science, to education, to medicine and social welfare. Furthermore, many sophisticated randomized designs have been developed to estimate various kinds of treatment effects under various settings of data generation. For example, within the category of uniform randomized experiment,[3] in addition to the traditional method of *completely randomized experiment*, where stratification is absent (i.e., $S = 1$ and S stands for number of strata), researchers have developed *randomized block experiments* where two or more strata are permissible (i.e., $S \geq 2$) and *paired randomized experiments* in which $n_S = 2$ (i.e., the number of study participants within stratum S is fixed at 2), $m_S = 1$ (i.e., the number of participants receiving treatment within stratum S is fixed at 1), and S could be reasonably large (Rosenbaum, 2002b).

A more important reason for studying randomized experiments is that statistical tests developed through randomized experiments may be performed virtually without assumptions, which is not the case for nonrandomized experiments. The class of randomization tests, as reviewed and summarized by Rosenbaum (2002b), includes

1. Tests for binary outcomes: *Fisher's* (1935/1971) *exact test,* the *Mantel-Haenszel* (1959) *statistic,* and *McNemar's* (1947) *test*
2. Tests for an outcome variable that is confined to a small number of values representing a numerical scoring of several ordered categories (i.e., an ordinal variable): *Mantel's* (1963) *extension of the Mantel-Haenszel test*
3. Tests for a single stratum $S = 1$, where the outcome variable may take many numerical values (i.e., an interval or ratio variable): *Wilcoxon's* (1945) *rank sum test*

4. Tests for an outcome variable that is ordinal and the number of strata S is large compared with sample size N: the *Hodges and Lehmann* (1962) *test using the signed rank statistic*

As opposed to drawing inferences using these tests in randomized designs, drawing inferences using these tests in nonrandomized experiments "requires assumptions that are not at all innocuous" (Rosenbaum, 2002b, p. 27).

1.3.3 Critiques of Social Experimentation

Although the randomized experiment has proven useful in many applications since Fisher's seminal work, the past three decades have witnessed a chorus of challenges to the fundamental assumptions embedded in the experimental approach. In particular, critics have been quick to note the complexities of applying randomized trials in studies conducted with humans rather than mechanical components, agricultural fields, or cups of tea. The dilemma presented in social and health sciences studies with human participants is that assigning participants to a control condition means potentially denying treatment or services to those participants; in many settings, such denial of services would be unethical or illegal. Although the original rationale for using a randomized experiment was the infeasibility of controlling covariates, our evaluation needs have returned to the point where covariant control or its variants (e.g., matching) becomes attractive. This is particularly true in social behavioral evaluations.

In a series of publications, Heckman and his colleagues (e.g., Heckman, 1979; Heckman & Smith, 1995) discussed the importance of directly modeling the process of assigning study participants to treatment conditions by using factors that influence participants' decisions regarding program participation. Heckman and his associates challenged the assumption that we can depend on randomization to create groups in which the treated and nontreated participants share the same characteristics under the condition of nontreatment. They questioned the fundamental assumption embedded in the classical experiment: that randomization removes selection bias.

Heckman and Smith (1995) in particular held that social behavioral evaluations need to explicitly address four questions, none of which can be handled suitably by the randomized experiment: (1) What are the effects of factors such as subsidies, advertising, local labor markets, family income, race, and gender on program application decisions? (2) What are the effects of bureaucratic performance standards, local labor markets, and individual characteristics on administrative decisions to accept applicants and place them in specific programs? (3) What are the effects of family background, subsidies, and local market conditions on decisions to drop out of a program and, alternatively, on the length of time required to complete a program? (4) What are the costs of various alternative treatments?

1.4 WHY AND WHEN A PROPENSITY SCORE ANALYSIS IS NEEDED

Drawing causal inferences in observational studies or studies without randomization is challenging, and it is this task that has motivated statisticians and econometricians to explore new analytic methods. The seven analytic models that we discuss in this book derive from this work. Although the models differ on the specific means employed, all seven models aim to accomplish data balancing when treatment assignment is nonignorable, to evaluate treatment effects using nonrandomized or nonexperimental approaches, and/or to reduce multidimensional covariates to a one-dimensional score called a *propensity score.* To provide a sense of why and when propensity score methods are needed, we use examples drawn from the literature across various disciplines. Propensity score analysis is suitable to data analysis and to causal inferences for a variety of studies. Most of these examples will be revisited throughout this book.

Example 1: Assessing the Impact of Catholic Versus Public Schools on Learning. A longstanding debate in education is whether Catholic schools (or private schools in general) are more effective than public schools in promoting learning. Obviously, a variety of selections are involved in the formation of "treatment" (i.e., entrance into Catholic schools). To name a few, *self-selection* is a process that lets those who choose to study in Catholic schools receive the treatment; *school selection* is a process that permits schools to select only those students who meet certain requirements, particularly minimum academic standards, to enter into the treatment; *financial selection* is a process that excludes from the treatment those students whose families cannot afford tuition; and *geographic selection* is a process that selects out (i.e., excludes) students who live in areas where no Catholic school exists. Ultimately, the debate on Catholic schools centers on whether differences observed in outcome data (i.e., academic achievement or graduation rates) between Catholic and public schools are attributable to the intervention or to the fact that the Catholic schools serve a different population. In other words, if the differences are attributable to the intervention, findings suggest that Catholic schools promote learning more effectively than do public schools, whereas if the differences are attributable to the population served by Catholic schools, findings would show that students currently enrolled in Catholic schools would always demonstrate better academic outcomes regardless of whether they attended private or public schools. It is infeasible to conduct a randomized experiment to answer these questions; however, observational data such as the National Educational Longitudinal Survey (NELS) data are available to researchers interested in this question.

Because observational data lack randomized assignment of participants into treatment conditions, researchers must employ statistical procedures to balance the data before assessing treatment effects. Indeed, numerous published studies have used the NELS data to address the question of Catholic school effectiveness; however, the findings have been contradictory. For instance, using propensity score matching and the NELS data, Morgan (2001) found that the Catholic school effect is the strongest only among those Catholic school students who, according to their observed characteristics, are least likely to attend Catholic schools. However, in a study that used the same NELS data but employed a new method that directly assessed selectivity bias, Altonji, Elder, and Taber (2005) found that attending a Catholic high school substantially increased a student's probability of graduating from high school and, more tentatively, attending college.

Example 2: Assessing the Impact of Poverty on Academic Achievement. Prior research has shown that exposure to poverty and participation in welfare programs have strong impacts on child development. In general, growing up in poverty adversely affects a child's life prospects, and the consequences become more severe with greater exposure to poverty (Duncan, Brooks-Gunn, Yeung, & Smith, 1998; Foster & Furstenberg, 1998, 1999; P. K. Smith & Yeung, 1998). Most prior inquiries in this field have applied a multivariate analysis (e.g., multiple regression or regression-type models) to samples of nationally representative data such as the Panel Study of Income Dynamics (PSID) or administrative data, although a few studies employed a correction method such as propensity score analysis (e.g., Yoshikawa, Maguson, Bos, & Hsueh, 2003). Using a multivariate approach with this type of data poses two fundamental problems. First, the bulk of the literature regarding the impact of poverty on children's academic achievement assumes a causal perspective (i.e., poverty is the cause of poor academic achievement), whereas the analysis using a regression model is, at best, correlational. In addition, a regression model or covariance control approach is less robust in handling endogeneity bias. Second, PSID is an observational survey without randomization and, therefore, researchers must take selection bias into consideration when employing PSID data to assess causal effects.

Guo and Lee (2008) have made several efforts to examine the impacts of poverty. First, using PSID data, propensity score models—including optimal propensity score matching, the treatment effects model, and the matching estimator—were used to estimate the impact of poverty. Second, Guo and Lee conducted a more thorough investigation of poverty. That is, in addition to conventional measures of poverty such as the ratio of income to poverty threshold, they examined 30 years of PSID data to create two new variables: (1) the number of years during a caregiver's childhood (i.e., ages 6–12 years) that a caregiver used Aid to Families With Dependent Children (AFDC) and (2) the percentage of time a child used AFDC between birth and 1997 (i.e., the time point when academic achievement data were compared). Last, Guo and Lee conducted both efficacy subset analysis and intent-to-treat analysis and compared findings. Results using these approaches were more revealing than previous studies.

Example 3: Assessing the Impact of a Waiver Demonstration Program. In 1996, the U.S. Congress approved the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA). Known as welfare reform, PRWORA ended entitlements to cash assistance that were available under the prior welfare policy, AFDC. As part of this initiative, the federal government launched the Waiver Demonstration program, which allowed participating states and counties to use discretionary funding for county-specific demonstration projects of welfare reform—as long as these demonstrations facilitated "cost neutrality." A key feature of the Waiver Demonstration program, as well as several other programs implemented under welfare reform, is that the county has the option of whether to participate in the Waiver Demonstration. Therefore, by definition, the intervention counties and comparison counties at the state level cannot be formed randomly. Counties that chose to participate differed from counties choosing not to participate. Evaluating such a nonrandomized program is daunting. Using a Monte Carlo study, Guo and Wildfire (2005) demonstrated that propensity score matching is a useful analytic approach for such data and an approach that provides less biased findings than does an analysis using the state-level population data.

Example 4: Assessing the Well-Being of Children Whose Parents Abuse Substances. A strong, positive association between parental substance abuse and involvement with the child welfare system has been established (e.g., English, Marshall, Brummel, & Coghan, 1998; U.S. Department of Health and Human Services, 1999). Substance abuse may lead to child maltreatment through several mechanisms, such as child neglect that occurs when substance-abusing parents give greater priority to their drug use than to caring for their children, or substance abuse can lead to extreme poverty and inability to provide for a child's basic needs (Magura & Laudet, 1996). Policy makers have long been concerned about the safety of children of substance-abusing parents. Drawing on a nationally representative sample from the National Survey of Child and Adolescent Well-Being (NSCAW), Guo et al. (2006) used a propensity score matching approach to address whether children whose caregivers received substance abuse services were more likely to have re-reports of maltreatment than were children whose caregivers did not use substance abuse services. Using the same NSCAW data, Guo et al. employed propensity score analysis with nonparametric regression to examine the relationship between the participation of a caregiver in substance abuse services and subsequent child outcomes; that is, they investigated whether children of caregivers who used substance abuse services exhibited more behavioral problems than did children of caregivers who did not use such services.

Example 5: Estimating the Impact of Multisystemic Therapy (MST). MST is a multifaceted, short-term (4–6 months), home- and community-based intervention for families with youths who have severe psychosocial and behavioral problems. Funding for MST in the United States rose from US$5 million in 1995, to approximately US$18 million in 2000, and to US$35 million in 2003. Most evaluations of the program used a randomized experiment approach, and most studies generally supported the efficacy of MST. However, a recent study using a systematic review approach (J. H. Littell, 2005) found different results. Among the problems observed in previous studies, two major concerns arose: (1) the variation in the implementation of MST and (2) the integrity of conducting randomized experiments. From a program evaluation perspective, this latter concern is a common problem in social behavioral evaluations: Randomization is often broken or compromised. Statistical approaches, such as propensity score matching, may be helpful when randomization fails or is impossible (Barth et al., 2007).

Example 6: Assessing Program Outcomes in Group-Randomized Trials. The Social and Character Development (SACD) program was jointly sponsored by the U.S. Department of Education (DOE) and the Centers for Disease Control and Prevention. The SACD intervention project was designed to assess the impact of schoolwide social and character development education in elementary schools. Using a scientific peer review process, seven proposals to implement SACD were chosen by the Institute of Education Sciences in the U.S. DOE, and the research groups associated with each of the seven proposals implemented different SACD programs in primary schools across the country. At each of the seven sites, schools were randomly assigned to receive either the intervention program or control curricula, and one cohort of students was followed from third grade (beginning in fall 2004) through fifth grade (ending in spring 2007). A total of 84 elementary schools were randomized to intervention and control at seven sites: Illinois (Chicago), New Jersey, New York (Buffalo, New York City, and Rochester), North Carolina, and Tennessee.

Evaluating programs generated by a group randomization design is often challenging, because the unit of analysis is a cluster—such as a school—and sample sizes are so small as to compromise randomization. At one of the seven sites, the investigators of SACD designed the Competency Support Program to use a group randomization design. The total number of schools participating in the study within a school district was determined in advance, and then schools were randomly assigned to treatment conditions within school districts; for each treated school, a school that best matched the treated school on academic yearly progress, percentage of minority students, and percentage of students receiving free or reduced-price lunch was selected as a control school (i.e., data collection only without receiving intervention). In North Carolina, over a 2-year period, this group randomization procedure resulted in a total of 14 schools (Cohort 1, 10 schools; Cohort 2, 4 schools) for the study: 7 received the Competency Support Program intervention, and 7 received routine curriculum. As it turned out—as is often the case when implementing randomized experiments in social behavioral sciences—the group randomization did not work out as planned. In some school districts, as few as four schools met the study criteria and were eligible for participation. Just by the luck of the draw (i.e., by random assignment), the two intervention schools differed systematically on covariates from the two control schools. Thus, when comparing data from the 10 schools, the investigators found the intervention schools differed from the control schools in significant ways: The intervention schools had lower academic achievement scores on statewide tests (Adequate Yearly Progress [AYP]), a higher percentage of students of color, a higher percentage of students receiving free or reduced-price lunches, and lower mean scores on behavioral composite scales at baseline. These differences were statistically significant at the .05 level using bivariate tests and logistic regression models. The researchers were confronted with the failure of randomization. Were these selection effects ignored, the evaluation findings would be biased. It is just at this intersection of design (i.e., failure of randomization) and data analysis that propensity score approaches become very helpful.

The preceding examples illustrate conditions under which researchers might consider propensity score modeling. The need for conducting propensity score analysis can also be determined by an imbalance check. This bivariate analysis of the equivalence of covariates by treatment condition may be viewed as an initial check of group or condition comparability. Balance checks help researchers discern whether data correction approaches more sophisticated than covariance control or regression modeling may be warranted. Because of its centrality in making analytic decisions, we present details of the imbalance check here.

As noted earlier, researchers are often concerned with the validity of inferences from observational studies, because, in such a setting, the data are generated by a nonrandom process; thus, to determine whether a study requires a correction other than simple covariance control, an initial test using a normalized difference score Δ_X may be undertaken (Imbens & Wooldridge, 2009). The test is basically a bivariate analysis using the treatment indicator variable and each covariate X, and X can be either a continuous or dichotomous variable. Δ_X is defined as follows:

$$\Delta_X = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{S_0^2 + S_1^2}}, \tag{1.1}$$

where $\bar{X}_1$ and $\bar{X}_0$ are the sample mean values of X, and S_1^2 and S_0^2 are the sample variances of X, for the treatment group and comparison group, respectively. Examples of applying Equation 1.1 to check normalized differences are shown in Section 6.5.2. Following Imbens and Wooldridge, a Δ_X exceeding .25 is an indication that selection bias exists and linear regression methods tend to be sensitive to the model specification. In other words, if an initial check of study data shows Δ_X exceeding .25 for numerous covariates, researchers should consider employing corrective approaches other than regression, or at least perform corrective analysis in conjunction with regression analysis. As a family of corrective approaches, propensity score models have promising properties and offer several advantages under the condition of imbalance.

1.5 COMPUTING SOFTWARE PACKAGES

At the time of the first edition of this book, few software packages offered comprehensive procedures to handle the statistical analyses described in subsequent chapters. Recently, however, more software programs have been developed for propensity score analysis. Our review of software packages indicates that Stata (StataCorp, 2007) and R (R Foundation for Statistical Computing, 2008) offer the most comprehensive computational facilities. Other packages, such as SAS, offer user-developed macros or procedures targeting specific problems (e.g., SAS Proc Assign may be used to implement optimal matching), but they do not offer the variety of analysis options that is available in Stata and R.

Table 1.1 lists the Stata and R procedures available for implementing the analyses described in this book. Just like the rapid growth of propensity score methods per se, computing programs have also developed at a fast pace. Table 1.1 shows some of the programs currently available. See also Stuart (2014), who provides a comprehensive list of software programs for implementing matching methods and propensity scores in R, Stata, SAS, and SPSS. We have chosen to use Stata to illustrate most approaches. We chose Stata based on our experience with the software and our conclusion that it is a convenient software package. Specifically, Stata's program ***test_condate*** can be used to test treatment effect heterogeneity described in Chapter 2; ***heckman*** and ***treatreg*** can be used to solve problems described in Chapter 4; ***psmatch2, pscore***, ***boost***, ***imbalance***, ***hodgesl, logistic, xtlogit***, ***xtmelogit***, and ***xtmixed*** can be used to solve problems described in Chapter 5; ***pscore, hte,*** and Stata programming commands can be used to solve problems described in Chapter 6; ***pweight*** function specified in a multivariate model can be used to solve problems described in Chapter 7; ***nnmatch*** can be used to solve problems described in Chapter 8; ***psmatch2*** can be used to solve problems described in Chapter 9; ***pweight*** function specified in a multivariate model and ***gpscore*** can be used to solve problems described in Chapter 10; and ***rbounds*** and ***mhbounds*** can be used to perform Rosenbaum's (2002b) sensitivity analysis described in Chapter 11. In each of these chapters, we will provide examples and an overview of Stata syntax. We provide illustrative examples for one R procedure (i.e., ***optmatch***), because this is the only procedure available for conducting optimal matching within R and Stata. All syntax files and illustrative data can be downloaded from this book's companion website (http://ssw.unc.edu/psa).

Many of the Stata programs described above were macros or ado files developed by users. At the time this second edition was completed, Stata released its version 13 (StataCorp, 2013).

This version of Stata for the first time includes a series of programs facilitating statistical analyses using propensity scores and other methods. Under the title of "***Treatment effects,***" this group of programs includes ***regression adjustment***, ***inverse-probability weights (IPW)***, ***doubly robust estimators***, ***matching estimators***, ***overlap plots***, and ***endogenous treatment estimators***. Many of these newly released programs offer functions similar to those described in this book, although the user-developed programs will continue to be needed for many analyses.

1.6 PLAN OF THE BOOK

Chapter 2 offers a conceptual framework for the development of scientific approaches to causal analysis, namely, the Neyman-Rubin counterfactual framework. In addition, the chapter reviews a closely related, and recently developed, framework that aims to guide scientific inquiry of causal inferences: the econometric model of causality (Heckman, 2005). The chapter includes a discussion of two fundamental assumptions embedded in nearly all outcome-oriented program evaluations: the ignorable treatment assignment assumption and the stable unit treatment value assumption (SUTVA). Violations of these assumptions pose challenges to the estimation of counterfactuals. The chapter provides a review of corrective methods other than propensity score analysis, particularly two methods that are widely employed in economic research, namely, the instrumental variables estimator and regression discontinuity design. The chapter offers a discussion on the importance of modeling treatment effect heterogeneity, two tests of effect heterogeneity, and an example to show the application of these tests.

Chapter 3 focuses on the issue of ignorable treatment assignment from the other side of the coin: strategies for data balancing when treatment effects can only be assessed in a nonexperimental design. This chapter aims to answer the key question of what kind of statistical methods should be considered to remedy the estimation of counterfactuals, when treatment assignment is not ignorable. Moreover, the chapter describes three closely related but methodologically distinctive approaches: *ordinary least squares* (OLS) regression, matching, and stratification. The discussion includes a comparison of estimated treatment effects of the three methods under five scenarios. These methods involve making simple corrections when assignment is not ignorable, and they serve as a starting point for discussing the data issues and features of more sophisticated approaches, such as the seven advanced models described later in the book. The chapter serves as a review of preliminary concepts that are a necessary foundation for learning more advanced approaches.

Chapters 4 through 10 present statistical theories using examples to illustrate each of the seven advanced models covered in this book. Chapter 4 describes and illustrates Heckman's sample selection model in its original version (i.e., the model aims to correct for sample selection) and the revised Heckman model developed to evaluate treatment effects. Chapter 5 describes propensity score matching, specifically the creation of matched samples using caliper (or Mahalanobis metric) matching and recently developed methods of optimal matching, propensity score matching with multilevel modeling, estimation of propensity scores with a generalized boosted regression, and various approaches for postmatching analysis of outcomes. Although Chapter 5 focuses on matching, sections on estimating propensity

Table 1.1 Stata and R Procedures by Analytic Methods

	Procedure Name and Useful References	
Chapter and Methods	*Stata*	*R*
Chapter 2		
Tests of treatment effect heterogeneity	***test_condate*** (Crump, Hotz, Imbens, & Mitnik, 2008)	
Chapter 4		
Heckman (1978, 1979) sample selection model	***heckman*** (StataCorp, 2003)	***sampleSelection*** (Toomet & Henningsen, 2008)
Maddala (1983) treatment effect model	***treatreg*** (StataCorp, 2003)	
Chapter 5		
Rosenbaum and Rubin's (1983) propensity score matching	***psmatch2*** (Leuven & Sianesi, 2003)	***cem*** (Dehejia & Wahba, 1999; Iacus, King, & Porro, 2008)
	pscore (Becker & Ichino, 2002)	***Matching*** (Sekhon, 2007)
		MatchIt (Ho, Imai, King, & Stuart, 2004)
		PSAgraphics (Helmreich & Pruzek, 2008)
		WhatIf (King & Zeng, 2006, 2007)
		USPS (Obenchain, 2007)
Generalized boosted regression	***boost*** (Schonlau, 2007)	***gbm*** (McCaffrey, Ridgeway, & Morral, 2004) ***twang*** (Ridgeway, McCaffrey, Morral, Griffin, & Burgette, 2013)
Optimal matching (Rosenbaum, 2002b)		***optmatch*** (Hansen, 2007)
Postmatching covariance imbalance check (Haviland, Nagin, & Rosenbaum, 2007)	***imbalance*** (Guo, 2008b)	
Hodges-Lehmann aligned-rank test after optimal matching (Haviland et al., 2007; Lehmann, 2006)	***hodgesl*** (Guo, 2008a)	

Chapter and Methods	*Procedure Name and Useful References*	
	Stata	*R*
Propensity score matching with multilevel data	***logistic, xtlogit, xtmelogit, xtmixed*** (StataCorp, 2007)	***glm, glmer, multilevel, nlme*** (R Foundation for Statistical Computing, 2013)
Chapter 6		
Propensity score subclassification	***pscore*** (Becker & Ichino, 2002)	***MatchIt*** (Ho, Imai, King, & Stuart, 2004) ***Zelig*** (Owen, Imai, King, & Lau, 2013).
	Use Stata programming commands to stratify the sample and then conduct aggregated analysis.	
Stratification-multilevel method	***hte*** (Jann, Brand, & Xie, 2010)	
Chapter 7		
Propensity score weighting	***pweight*** function specified in a multivariate model (StataCorp, 2003)	***probs*** or ***weight*** function used in ***svydesign*** (R Foundation for Statistical Computing, 2013) ***twang*** (Ridgeway et al., 2013)
Chapter 8		
Matching estimators (Abadie & Imbens, 2002, 2006)	***nnmatch*** (Abadie, Drukker, Herr, & Imbens, 2004)	***Matching*** (Sekhon, 2007)
Chapter 9		
Kernel-based matching (Heckman, Ichimura, & Todd, 1997, 1998)	***psmatch2*** (Leuven & Sianesi, 2003)	
Chapter 10		
Propensity score analysis of categorical or continuous treatments	***pweight*** function specified in a multivariate model (StataCorp, 2003), ***gpscore*** (Bia & Mattei, 2008)	
Chapter 11		
Rosenbaum's (2002b) sensitivity analysis	***rbounds*** (Gangl, 2007), ***mhbounds*** (Becker & Caliendo, 2007)	***rbounds*** (Keele, 2008)

scores and strategies for developing optimal models serve also as a guide for the methods described in Chapters 6, 7, 9, and 10. In these chapters, virtually the same approaches are used to estimate propensity scores. Chapter 6 focuses on propensity score subclassification, a method that can be applied to outcome variables that are not normally distributed and special types of models such as structural equation modeling. Chapter 7 describes propensity score weighting, a robust approach that can also be applied to various types of outcome variables, such as time-to-event data, and complex outcome analyses using structural equation modeling. Chapter 8 describes a collection of matching estimators developed by Abadie and Imbens (2002), who provide an extension of Mahalanobis metric matching. Among the attractive features of this procedure is its provision of standard errors for various treatment effects. Chapter 9 describes propensity score analysis with nonparametric regression. Specifically, it describes the two-time-period difference-in-differences approach developed by Heckman and his colleagues (Heckman, Ichimura, & Todd, 1997, 1998). Chapter 10 describes methods to model doses of treatment. This chapter extends the basic methods for binary treatment conditions (treated and control) to more complex situations in which a treatment variable has more than two conditions and can be either categorical or continuous.

Chapter 11 reviews selection bias, which is the core problem all statistical methods described in this book aim to resolve. This chapter gives the selection bias problem a more rigorous treatment: We simulate two settings of data generation (i.e., selection on observables and selection on unobservables) and compare the performance of six models under these settings using Monte Carlo studies. Hidden selection bias is a problem that fundamentally distinguishes observational studies from randomized experiments. When key variables are missing, researchers inevitably stand on thin ice when drawing inferences about causal effects in observational studies. However, Rosenbaum's (2002b) sensitivity analysis, which is illustrated in Chapter 11, is a useful tool for testing the sensitivity of study findings to hidden selection. This chapter reviews assumptions for all seven models and demonstrates practical strategies for model comparison.

Finally, Chapter 12 focuses on continuing issues and challenges in the field. It reviews debates on whether propensity score analysis can be employed as a replacement for randomized experiments. It comments on recent advances. And it suggests directions for the development of new approaches to observational studies.

NOTES

1. Excel can be used to calculate the number of combinations of 8 things taken 4 at a time by typing the following in a cell: =COMBIN(8,4), and Excel returns the number 70.
2. If the tea taster gives an outcome other than 11110000, then the number of assignments having six exact agreements remains 17. However, there will be a different set of 17 assignments than those presented here.
3. Uniform here refers to equal probability for elements in the study population to receive treatment.