

## CHAPTER 1. BASIC CONCEPTS OF PROPENSITY SCORE METHODS

In behavioral and social sciences, due to practical or ethical barriers, researchers often cannot collect data from random trials (Bai, 2011). Therefore, observational studies are often used to make causal inferences (Pan & Bai, 2015a; Shadish, Cook, & Campbell, 2002). Unfortunately, selection bias in observational research often poses a threat to the validity of these studies (Rosenbaum & Rubin, 1983). Selection bias occurs when the participants in one study condition (e.g., the treatment group) are systematically different in their preexisting characteristics from those in another condition (e.g., the control group). For example, if participants self-select into a treatment group, they may be more motivated, more conscientious, or more ambitious than those in the control group. When participants are randomly assigned to groups, this bias is usually reduced. On expectation, participants who are randomly assigned will have similar distributions of characteristics between the groups (i.e., those in the control group are just as motivated, conscientious, and ambitious as those in the treatment group). When covariates are equivalent across groups, they are balanced, and researchers can reasonably infer that any differences between the groups on the outcome variable are due to the causal (predictor or independent) variable. If covariates are not balanced, as is often the case in observation studies, the preexisting differences between the groups may be responsible for any differences that we see in the outcome variables, resulting in a spurious treatment effect. To increase the validity of the treatment effect estimation, a variety of statistical adjustments may be used to reduce selection bias; however, some are more effective than others.

Over the past decades, propensity score (PS) methods have become increasingly popular for improving the validity of causal studies, as they can produce results that mimic those from true experimental designs when used appropriately (Rosenbaum & Rubin, 1985). Since their introduction by Rosenbaum and Rubin in 1983, PS methods have been used in many fields, such as education (e.g., Clark & Cundiff, 2011; Guill, Lütcke, & Köller, 2017; Hong & Raudenbush, 2005), epidemiology (e.g., Austin, 2009; Thanh & Rapoport, 2016), psychology (e.g., Gunter & Daly, 2012; Kirchmann et al., 2012), economics (e.g., Baycan, 2016; Dehejia & Wahba, 2002), political science (Seawright & Gerring, 2008), and program evaluation (e.g., Duwe, 2015). For example, Gunter and Daly used propensity scores when examining the relationship between violent video games and

deviant behavior. They found that after accounting for self-selection of the type of games played, PS matching decreased the treatment estimates, indicating that video games have a weaker effect on violent and deviant behaviors than previous research had suggested. Guill et al. compared several PS models to account for selection bias when examining how students on an academic track differed from those either on a nonacademic track or attending a comprehensive school on cognitive development. Duwe used PS matching to evaluate how well a prisoner reentry program reduced recidivism and increased postrelease employment.

Although research across a variety of fields has demonstrated that PS methods consistently improve the accuracy of treatment effects, there are still some challenges for researchers who apply these methods to their own empirical studies (Pan & Bai, 2016). While there are other volumes that cover more specific problems in greater detail than this text (e.g., Guo & Fraser, 2015; Leite, 2017; Pan & Bai, 2015a), this book provides an introduction to the general use and practical applications of PS methods so that after reading this book, the reader should be able to meet the following goals:

1. Understand when it is or is not appropriate to use PS methods, given a researcher's goal, design, and available data;
2. Be able to assess the common support of the estimated propensity scores (i.e., how well the propensity scores are similar across groups);
3. Be able to model and estimate propensity scores that will sufficiently account for selection bias in an observational study;
4. Be familiar with the most common PS methods (i.e., PS matching, subclassification, inverse probability weighting, covariate adjustment, and doubly robust adjustments) and have a sense of how to select the most appropriate method based on their research designs, data, and propensity scores;
5. Know how to use those PS methods;
6. Know how to assess individual covariates for balance across groups;
7. Know how to estimate their adjusted treatment effect;
8. Understand the limitations when using PS methods; and
9. Get to know a variety of software packages used to implement PS methods through the book's website.

The book is structured in this order so that the reader can follow all the steps necessary to complete a PS procedure. Each chapter is devoted to one

or two of these goals as listed above. Chapter 1 introduces the basic concepts of making causal inferences from experimental and observational studies, and then discusses propensity scores in terms of *what* they are, *when* to use them, *why* we use them (Goal 1), and the assumptions that need to be met when using them (Goal 2). Chapter 2 focuses on how to select appropriate covariates and model propensity scores (Goal 3). Chapter 3 discusses four commonly used PS methods (matching, stratification, weighting, and covariate adjustment) (Goals 4 and 5). Chapter 4 covers how to evaluate the balance of covariate distributions, how to estimate the adjusted treatment effect, and how robust the treatment effect estimation is against hidden bias (Goals 6 and 7). Chapter 5 summarizes the key points of PS methods, provides some general guidelines for handling common problems with PS methods, and introduces some new developments in PS methods (Goal 8). Finally, the companion website for this book at [study.sagepub.com/researchmethods/qass/bai&clark](http://study.sagepub.com/researchmethods/qass/bai&clark) provides instructions, code, and interpretations of output for a variety of statistical software packages that are commonly used to implement PS methods (Goal 9).

To help readers better understand the procedures for implementing PS methods, Chapters 2, 3, and 4 include an example of how to apply these procedures to real world data. These examples demonstrate each step of the PS procedures that correspond to what we previously discussed in the chapter. The data are a subset of a dataset that is publicly available from the Inter-University Consortium for Political and Social Research (ICPSR 35683). The data were originally used to assess the Playworks intervention, which is a recess program for elementary school children intended to improve social and emotional skills by teaching safe, engaging forms of play ([www.playworks.org](http://www.playworks.org)). Because software packages often change, the dataset, program codes, outputs, and interpretations of the output are provided on the book's website. Readers are encouraged to replicate the examples using our code and check their results with those we provide online.

## 1.1 Causal Inference

### 1.1a Experimental Design and Observational Studies

In an experimental design, it is assumed that we obtain treatment and control groups with equal distributions of group member characteristics (except for the treatment condition) through random sampling and random assignment, thereby limiting potential selection bias, so that the factor of interest is the only cause of an effect. In contrast, researchers conducting an observational study, from which conclusions are drawn based on results from data collected without random assignment, are less confident when

making causal inferences. To better understand why this is an issue, in this section we briefly discuss the basic concept of causal inference and illustrate the importance of good research designs.

Suppose that we are interested in studying whether a recess program impacts social skills among elementary school children. According to the counterfactual framework for modeling causal effects, the true treatment effect for each child would be the difference between the treated outcome and the counterfactual (i.e., the outcome in the absence of the treatment) (Holland, 1986; Rubin, 1974). In this context, we would need to compare the social skills of each child who participated in the recess program (the “treatment” participants) to *the same child’s* social skills if he or she had not been in the recess program (the counterfactual).

Obviously, we cannot observe social skills under both of these conditions at the same time, since children cannot simultaneously be in the program and not be in the program. Therefore, as a reasonable alternative, one can estimate the *average treatment effect* (ATE) (Holland, 1986; Rubin, 1974; Winship & Morgan, 1999) for the population. To assess the ATE for the children’s social skills, we examine the difference between the expected value of social skills for all the children in the recess program and the expected value of social skills for all the children who were not in the recess program. If we randomly select students from the population and randomly assign them into the recess program, the ATE is an unbiased estimate of the treatment effect because the recess (treatment) group does not, on average, differ systematically from those who were not in the recess program (comparison group) on their observed and unobserved background characteristics.

However, in many research situations, randomized control trials (RCTs) or true experiments, in which participants are randomly selected and assigned to groups, are not always feasible. In some research situations, it is not possible to randomly assign participants to conditions, and in others it is not ethical to randomly assign them. For example, it is highly unlikely that we have the ability to manipulate parents’ expectations, force people to seek therapy, or control who attends college. Even when random assignment is possible, it may not be ethical to randomly assign participants to risky conditions, such as smoking, alcohol use, cancer, sexually transmitted diseases, child abuse, or homelessness. However, the absence of random assignment should not prevent us from studying how psychotherapy affects depression (Bernstein et al., 2016); how alcohol use affects coronary heart disease (e.g., Fillmore, Kerr, Stockwell, Chikritzhs, & Bostrom, 2006); how maternal smoking influences birth weight and preterm birth (e.g., Ko et al., 2014); or how types of child abuse (physical, sexual, or emotional) impact depression and aggression in its victims (e.g., Vachon, Krueger, Rogosch, & Cicchetti, 2015).

For example, when studying how parents' expectations of their children's academic success influence mathematics achievement, we cannot assign students to parents with high or low expectations, nor can we manipulate parents' expectations. Therefore, it is very likely that students' background characteristics in the two groups are significantly different, which may also influence their math achievement scores. Knowing that students are different on their characteristics other than just their parents' expectations, we cannot directly assess the impact of parents' expectations on students' math achievement using the observational data without controlling for other influential factors. The unbalanced distributions of the influential factors (often called confounding variables or covariates) between the two groups create selection bias, which usually causes a biased ATE. Naturally, our next question is, how can we draw *valid*, causal conclusions from observational studies? The next section addresses how this can be achieved.

### *1.1b Internal Validity of Observational Studies*

A statistical causal inference is a claim made about a cause-and-effect relationship between two or more variables from a statistical model. Therefore, the validity of a statistical causal inference, also called internal validity (Shadish et al., 2002), refers to a researcher making a *reasonable* inference from a statistical analysis of the data in which there is little doubt that a causal relationship exists. Selection bias is a considerable threat to the validity of statistical causal inference in observational studies. As we discussed in the previous section, selection bias refers to systematic differences in distributions of covariates that result in incomparable groups (e.g., people in the treatment group are older, more motivated, or more educated than those in the comparison group). Selection bias typically occurs when observed (measured) covariates or hidden (unmeasured) covariates are not accounted for in statistical models or controlled for in the design, which results in spurious estimates of causal effects (Rosenbaum, 2010). For instance, using the previous example of parents' expectations on students' academic performance, existing literature indicates that students' gender is related to both students' math achievement (Fennema & Sherman, 1997) and parents' expectations. Therefore, gender may be a confounding variable that influences students' math achievement because students cannot be randomly assigned to parents with high expectations or low expectations. In this case, we cannot make any *valid* causal claims regarding the impact of parents' expectations on students' math achievement without controlling for the influence from the confounding factors. Moreover, students' achievement in mathematics is also related to students' personal beliefs (Gutman, 2006; Schommer-Aitkins, Duell, & Hutter, 2005); their peers'

influence (Hanushek, Kain, Markman, & Rivkin, 2003); their reading abilities (Hill, Rowan, & Ball, 2005); environmental variables (Koth, Bradshaw, & Leaf, 2008); sociodemographic variables (e.g., ethnicity, socioeconomic status); and school compositions (Entwisle & Alexander, 1992), which may also confound the effect of parents' expectations on student achievement. With many such confounding variables, it is highly unlikely that all of the covariates in the study would be balanced between the high and low expectation groups. If the distributions of these covariates are not balanced, any estimates made without accounting for this imbalance would weaken the validity of the statistical causal inference of the study.

From the above example, it is clear that we cannot directly analyze observational data for causal effect without adjusting or controlling for the confounding variables. The confounding variables can be hidden (not measured), nonmeasurable, or observable (measured and available to the researcher). If these variables are *observable*, it is possible to reduce selection bias and improve the validity of the statistical causal inference by adjusting or controlling for those covariates.

### *1.1c Existing Methods to Reduce Selection Bias*

In many cases, we are not able to randomly select and assign participants to grouping conditions due to the constraints of the specific independent variable (e.g., researchers cannot randomly assign biological sex) or the will of the participants (e.g., participants are more likely to choose or be required to enter a drug rehabilitation program than to be randomly assigned into it). Therefore, we must find some way of balancing non-equivalent groups to increase the validity of causal inference when randomized trials are not feasible. Several approaches that are commonly used to control the influence of covariates and confounding factors are to (a) use designs that test or rule out alternative causal explanations, (b) use designs that balance groups on specific covariates, (c) account for known sources of bias (observed covariates) through statistical models that adjust the treatment effects, and (d) combine two or more of these approaches (Shadish et al., 2002).

The first option is achieved by adding design elements, which are variables or conditions added to a research design to assess threats to validity by varying experimental conditions. These commonly include comparison groups (e.g., control, placebo, partial treatment) or observations over time (e.g., pretests, follow-up measures). For example, adding a sugar pill as a placebo to a medical experiment may help researchers determine whether observed effects are due to the active ingredients in a medication or a patient's belief that the treatment will be effective. Adding a pretest

(even when participants are randomly assigned to groups) is common in educational studies, as it allows researchers to examine the difference in learning outcomes after instruction between two or more teaching methods or programs while controlling for preexisting characteristics that may influence students' performance.

While researchers agree that adding relevant elements to quasi-experimental designs can be effective in reducing threats to internal validity (Larzelere & Cox, 2013; Murnane & Willett, 2011; Shadish et al., 2002), this approach often requires a considerable amount of advance planning, complex statistical analyses, and available participants. Furthermore, these designs may not be feasible to carry out as a randomized study if assignment to conditions could not be controlled by the researcher.

An instrumental variable (IV) model is another control method that uses a variable that is correlated with the predictor (or causal variable) but is not associated with the change in the outcome variable. An IV can be correlated with the outcome variable, but it must not explain the change in the outcome variable. For example, when attempting to estimate the causal effect of parents' expectations on student achievement in mathematics, the correlation between parents' expectations and students' math scores does not imply that parents' expectations can cause students' math scores to change. Other variables may affect both parents' expectations and student achievement, or student achievement may affect parents' expectations. Since we cannot manipulate parents' expectations of their children, we may estimate the causal effect of parents' expectations on student achievement by using parents' income as an instrument. This assumes that parents' income impacts their expectations, but income is only correlated with student achievement through the effect of parents' expectations. If we find that parents' income and student achievement are correlated, this may be evidence that parents' expectations have a causal effect on student achievement. Unfortunately, despite Bowden and Turkington's (1990) claim that IV models produce results comparable to an experimental design, in practice, it can be difficult to correctly identify appropriate IVs to produce consistent treatment effect estimates (Land & Felson, 1978).

Researchers can also match participants on one or several potentially confounding variables either before or after an intervention to achieve similarity between treatment and control groups (Rubin, 2006). This is usually used in quasi-experimental designs. Although either continuous (e.g., age or parents' income) or categorical (e.g., gender or ethnicity) variables may be used in this matching process, it is easier to match on a couple of categorical variables than on several variables. Despite the common use of this method in quasi-experimental design, traditional matching presents two problems: (a) It is difficult to find exact matches for continuous variables,



and (b) it is difficult to match group members on multiple covariates, even with categorical variables. Using the parents' expectation study as an example, if we wanted to match on parents' income, this would require that we find the same parental income (e.g., \$65,000) for a child in the treatment group and in the control group. Given the variability of incomes, it is unlikely that we would find many parents in each group that have the same income. Matching on a single categorical variable, such as gender, would not be difficult; however, finding a child in the control group with the same gender, ethnicity, native language, and family composition as each child in the treatment group would limit the number of potential matches. By limiting the number of matches made between the treatment and control groups, the sample size is reduced, which also decreases statistical power and generalizability of research results.

The first problem can be resolved by using proximal matching, which matches members based on *similar* values (e.g., a student whose parents' income is \$65,000 can be matched with a student whose parents' income is \$64,800), rather than exact values. The second problem *could* be reduced by limiting the number of matching variables to one or two. However, this would also restrict the number of confounding variables that are balanced; therefore, estimated treatment effects are still biased even after matching.

Another common strategy is to control the confounding factors in non-randomized studies by using traditional covariate adjustment, such as analysis of covariance (ANCOVA) or a form of regression (e.g., ordinary least squares or logistic). These approaches partial out the effects of confounding variables on the treatment effect by including covariates in the statistical model (Eisenberg, Downs, & Golberstein, 2012; Jamelske, 2009; Ngai, Chan, & Ip, 2009). In the simplest case, researchers may use a pretest observation as a covariate with the hope of controlling for the group differences on pretest scores. More commonly, researchers will include several other confounding variables as covariates, knowing that treatment groups are probably different on those variables that also influence outcome estimations other than just pretest scores. Even though traditional covariate analyses *can* control for confounding factors to some extent (Leow, Wen, & Korfmacher, 2015; Stürmer et al., 2006), using these approaches presents some theoretical and practical problems. First, these statistical models may be easily misspecified due to small or unequal sample sizes, violations of statistical assumptions, or covariates that may not be able to sufficiently account for confounding due to the limited number of covariates that can be included in a specific model or an unmeasured confounding variable. While adding additional covariates may reduce confounding, each new covariate added to the model will reduce the statistical power.



A second major problem in using traditional covariate adjustments is that these analyses do not directly model bias. That is, covariates are not weighted according to how well they balance covariates, but rather how well they relate to the dependent variable. Therefore, rather than accounting for differences between groups in the covariate, they focus on accounting for the shared variance between the covariate (for all participants) and the outcome variable. For example, in a job training program, if the correlation between the starting and posttest salaries was high (e.g.,  $r = .7$ ), the covariate alone would explain 49% of the variance of the posttest salary. While this would still leave 51% of the variance unexplained, the unique contribution of the job training program may not be strong enough to be detected as a significant effect through an ANCOVA model. Despite the popularity of using traditional covariate adjustments to account for selection bias, they may not be appropriate statistical procedures for reducing selection bias.

Another significant limitation of covariance analysis is that including several covariates in the model simultaneously may reduce statistical power. However, if researchers limit the number of covariates, they may fail to control for all influential factors and still end up with a biased estimate of the causal effect. For example, in our job training example, there are many factors related to salary increases, so if we account for only some of these factors, the effect of job training on salary increase may not be estimated correctly. Therefore, only in some cases will covariance analysis be effective in controlling for the confounding factors from selection bias. Thus, it is clear that we need better methods. Although there are a variety of procedures that suitably model and reduce selection bias in observational studies (e.g., Camillo & D'Attoma, 2010; Heckman, 1979), some of the most widely used approaches use propensity scores (Rosenbaum & Rubin, 1983). The following sections will focus on the basic concepts related to this approach.

## 1.2 Propensity Scores

### 1.2a *What Is a Propensity Score?*

A propensity score is the probability that a participant would be assigned to a particular study group based on a set of covariates (Rosenbaum & Rubin, 1983). Most often, propensity scores are estimated as the likelihood that a person would be assigned or self-select into a treatment condition (see Chapter 2 for details as to how they are computed). As probabilities, propensity scores range from 0 to 1. Scores above .5 predict that a participant will be in the treatment group, and those below .5 predict that a participant will be in the control or comparison group. However, the goal of

propensity scores is not to perfectly predict assignment condition, but to create a single composite score to represent the whole set of covariates that can be used to account for group differences on all observed characteristics or confounding factors due to selection. This also assumes that participants with the same propensity scores will have the same distributions of observed covariates between the treatment and comparison groups. As such, the propensity score can then be used with a variety of statistical adjustments that should make the background characteristics or covariates of the participants in the treatment group comparable to those in the control or comparison group—as one would see with random assignment (Rosenbaum & Rubin, 1983). Common statistical adjustments used in PS methods include (a) matching, which pairs participants from treatment and control groups based on the proximity of their propensity scores; (b) subclassification (or stratification), which groups participants who are matched on several strata based on their propensity scores; (c) weighting, which multiplies outcome observations by a weight based on the propensity score; and (d) covariate adjustment, which uses propensity scores as a covariate in an ANCOVA or regression. These adjustment methods and how to conduct them are described more fully in Chapter 3. In theory, PS methods should balance the treatment groups on all of the observed covariates used to compute the propensity scores and reduce the bias caused by nonrandom assignment. If propensity scores are modeled appropriately, the adjusted treatment effects should be unbiased (Rosenbaum & Rubin, 1985).

### *1.2b Why Use Propensity Scores?*

PS methods may not be our first choice in controlling for bias in research, but they may be the best alternative to random assignment, as they address selection bias at the design level, as opposed to other statistical control procedures. As discussed in Section 1.1c, several existing methods can be used to control for confounding variables in observational studies. Under certain conditions, these methods can be effective in reducing bias. However, they also have several limitations, many of which PS methods can reduce. Like instrumental variables, covariate matching, and covariate adjustments, PS methods can also be conducted on existing data. Therefore, they permit the use of archival data to balance nonequivalent groups when designs cannot be altered.

While both the instrumental variable approach and covariate matching can reduce bias, these procedures only allow researchers to balance groups on the variables included in these adjustments. In many cases, only a single variable is used as the instrumental variable, which needs to

meet certain conditions that can be difficult to operate or identify (e.g., it must correlate with the treatment variable, but not the change in the outcome variable). Since it is probable that selection bias is affected by several variables, not all of these would be equally distributed between the treatment and control groups. Therefore, even if the instrumental variable meets the conditions for a certain analysis, selection bias may not be sufficiently reduced.

When matching on multiple covariates, it is very difficult to match on all of them simultaneously, as each additional covariate limits the number of viable matches. This often means that researchers must either match on several variables with limited levels (e.g., biological sex with options for only male or female, or age with options for only young or old) or select only a few influential variables with several levels (e.g., high school GPA or ACT when participants self-select into college). A better solution would be to use a composite score that aggregates several variables into one.

As a composite score, a propensity score combines the simplicity and statistical power of using a single score with the thoroughness of using multiple covariates by accounting for the variance of several variables concurrently (Rosenbaum & Rubin, 1983). Propensity scores aggregate multiple covariates into a single score, and covariates are weighted in a way that considers their relative importance in assignment to conditions. This solves the problems presented not just when using instrumental variables and covariate matching, but when using traditional covariate adjustment too.

Although traditional covariate adjustment can accommodate several covariates, statistical power can still be affected when trying to include several covariates, especially when using a small sample size. More importantly, propensity scores actually model selection bias, not the predictability of the dependent variable. Therefore, by using PS methods, researchers can actually account for statistical estimation *bias* from model misspecification due to design issues, rather than how the individual covariates relate to the outcome variable. This is why matching, stratifying, or statistical adjustments using propensity scores often reduce selection bias better than analysis of covariance or multivariate models (Grunwald & Mayhew, 2008; Peterson et al., 2003).

Despite the advantages that PS methods have over other methods used to reduce selection bias, propensity scores still have their limitations. Several conditions and assumptions should be met when using propensity scores, which are discussed in the next few sections. Like most statistics, if these assumptions are not met, propensity scores may not effectively reduce selection bias. These limitations and ways of addressing them are discussed in greater detail in Chapter 5.

### *1.2c When to Use Propensity Scores*

PS methods have been used to reduce group selection bias or adjust treatment effects in nonrandomized experiments in a variety of behavioral and social science fields (Baycan, 2016; Gunter & Daly, 2012; Kirchmann et al., 2012), and their use has increased exponentially within the past few decades (Bai, 2011). Unfortunately, their increased popularity could also lead to misuse (Pan & Bai, 2016). Like most statistical methods, they are appropriate only under certain conditions. PS methods are intended to balance group data when treatment assignment is nonignorable (e.g., assignment is not random, clearly specified, or maintained by participants); assess treatment effects when using quasi-experiments or other types of group comparisons using observational data; and aggregate several covariates into a single variable (the propensity score) to be used for statistical adjustments (Guo & Fraser, 2015; Shadish, 2010).

Because PS methods were created to improve internal validity, they should be used when researchers attempt to draw causal inferences from their observational studies. Propensity scores are used to account for preexisting individual characteristics that may be related to the treatment conditions tested for causal effects; therefore, we must be able to establish that the intended cause (even if it isn't a treatment or intervention) precedes the effect.

While propensity scores can be applied to a variety of nonrandomized experiments, they are intended to test causal effects from observational studies in which the assignment method is unknown. This may include quasi-experiments, natural experiments, or causal comparative studies. Within these studies, there are several ways in which assignment can be nonrandom, but corrected with PS methods:

1. Participants may have self-selected into a treatment. For example, when examining how the mode of instruction affects academic performance in college students, students may choose to sign up for an online course (treatment) instead of a face-to-face course (comparison) because it fits their schedules.
2. Someone assigned participants to groups based on an inconsistent or unknown criterion. If more than one person is determining who gets into the treatment, each person may use different standards for admission, or administrators may make exceptions for some by altering the criterion. For example, some children may be admitted to a gifted education program simply because they have ability scores that exceed 130, while others (who demonstrate high motivation to success or independence) are admitted with scores of 120.
3. The causal variable is not directly manipulated by the researcher. In cases of causal comparative or natural experiments, the event or

characteristic that we assume is causal is not a treatment or intervention that is imposed by a researcher; it is an existing characteristic or accidental event. Examples of these may include biological sex, birth order, marital status, socioeconomic status, and medical condition. A more specific example is illustrated in Almond's (2006) study that examined the effects of prenatal exposure to influenza on long-term health, education, and economic outcomes.

In all of these examples, the reasons for participants' assignment to treatment conditions is unknown or unclear, and using PS methods would be appropriate. However, if the assignment is based on a known (and maintained) criterion, such as when alcoholics are assigned to substance abuse programs based on the severity of their addictions, a regression discontinuity design (RDD) may be more effective and easier to use than PS methods. In theory, an RDD works under the same principles as a randomized control trial, in that we know the selection mechanism; therefore, we can control for it. By assigning participants to groups based on the value of a baseline characteristic, this serves as a proxy to random assignment and should account for selection bias. According to Shadish (2010), "Such assignment is called ignorable because potential outcomes are unrelated to treatment assignment once those known variables are included in the model, so an unbiased estimate can still be obtained" (p. 6). However, this assumes that the criterion for assignment is strictly followed, and that if participants are assigned to groups based on more than one variable, all assignment variables are included in the statistical model.

Finally, to balance covariates on several characteristics, researchers must have several measured covariates, which are related to both selection into conditions and the outcome variable, available to include in PS models. If researchers conduct their study using secondary data that is limited to a few demographic covariates, it is unlikely that they will be able to sufficiently model the selection process. In such cases, propensity score methods may not sufficiently reduce bias (Steiner, Cook, Shadish, & Clark, 2010). Therefore, it is recommended that researchers consider what variables are likely to influence assignment to conditions *before* data are collected so that these can be measured or use existing data with sufficient covariates.

## 1.3 Assumptions

### 1.3a The Ignorable Treatment Assignment Assumption

One of the assumptions when using PS methods is that assignment to treatment conditions is independent of the treatment effect after accounting for a set of observed covariates. In a randomized experiment, this assumption is

often met even without accounting for covariates, since (on expectation) random assignment balances all covariates between treatment conditions. Of course, this assumption is not guaranteed in a quasi-experiment, particularly when participants self-select into conditions. Under this assumption, if the distributions of the propensity scores are balanced between the treatment conditions, the distributions of the covariates used for obtaining propensity scores are also equal between the treatment conditions. Therefore, we assume that selection bias has been eliminated (or sufficiently reduced) after making statistical adjustments with the propensity scores, provided that all the confounding variables are measured. This is why we use PS methods in the first place.

One way to verify that selection bias has been reduced after using PS adjustments is to examine the relationship between treatment conditions and each observed covariate. A difference between the group means (or proportions when covariates are categorical) suggests that the covariate is unbalanced and violates this assumption. Chapter 4 more fully describes various methods for testing the balance of covariates.

Of course, we can only test for covariate balance on the variables that we measured and included in the PS estimation model. Although researchers should attempt to control for all reasonable sources of bias in the set of observed covariates used to estimate the propensity scores, it is likely that some unmeasured or unobserved covariates are not included; thus, selection bias remains even after PS adjustments. In this case, these omitted variables are sources of hidden bias that still affect treatment effects.

For example, if a covariate, such as risk for child abuse, is related to treatment assignment and the outcome, but is not included in the PS estimation, the treatment effects will still be biased. When the propensity scores calculated from the set of covariates do not represent all influential covariates, they cannot balance the distributions of all covariates between the groups. In this case, the ignorable treatment assignment assumption is not met using PS methods. When covariates were limited to only a few commonly available demographic covariates (e.g., age, ethnicity, sex, and marital status), less than half of the selection bias was removed (Steiner et al., 2010). Therefore, it is essential that all the covariates that contribute to selection bias are included in the PS model. Chapter 2 provides more guidance on how to select covariates so that this assumption is met.

### *1.3b The Stable Unit Treatment Value Assumption*

The second assumption when using PS methods is that the treatment effect for each individual will not depend on how each person gets into his or her respective condition. This requires that (a) the outcome does not depend on the assignment procedure (i.e., randomized or self-selection) and

(b) the treatment is the same for all participants in the treatment group (Holmes, 2014; Rosenbaum & Rubin, 1983). According to Cox (1958), “The observation of one unit should be unaffected by the particular assignment of treatment to the other units” (p. 19). When implementing PS methods, such as PS matching, the stable unit treatment value assumption (SUTVA) assumes that (a) within a matched pair, Participant A in the treatment group and Participant B in the control group have the same likelihood of being assigned to the treatment or control group; and (b) Participant A receives the same type and amount of treatment as the other participants in the treatment group who were selected through PS matching.

SUTVA is violated when the outcome depends on the version of the treatment participants receive or when there is an interaction between participants that would allow them to share the treatment. This can also be explained in terms of specific threats to validity: (a) unreliability of treatment implementation, in which treatment is not given consistently to each person in the treatment condition; (b) compensatory equalization, when participants in the control group receive an alternate version of the treatment; (c) compensatory rivalry, when participants in the control group are motivated to perform as well on the outcome as those in the treatment group; (d) resentful demoralization, when participants in the control group reduce their effort on the outcome because they did not receive the treatment; and (e) treatment diffusion, when the participants in the control group learn the treatment for those in the treatment group (Shadish et al., 2002). Under these circumstances, the participants did not receive the treatment (or lack of treatment) that they were assigned to receive. Clearly, we cannot make reasonable inferences about the effect of treatment if a participant actually receives a different treatment.

### *1.3c Sufficient Common Support or Overlap*

The third assumption implies that there is sufficient overlap in the distributions of the propensity scores estimated for the treatment and control groups; that is, the two groups being compared share a common support region of propensity scores in the sample data. This presumes that participants with the same propensity scores have an equal chance of being in either the treatment or control group based on the similarity of their background characteristics or covariates, which would allow us to isolate the treatment and make a reasonable (unbiased) comparison between the two groups. For example, if two employees both have a propensity score of .7, each has a 70% chance of being in a job training program based on his or her background characteristics. Then, we can reasonably compare their salaries after one completes the training program while the other does not.



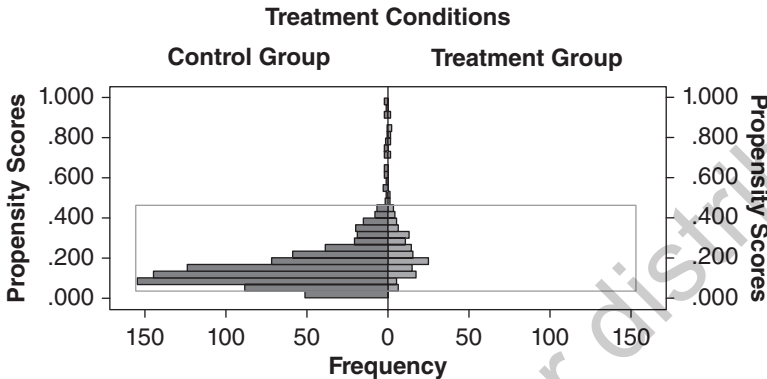
If most of the members in the treatment group have propensity scores that are similar to those in the control group, we assume that the two groups are comparable. The proportion of the similarity of the propensity scores of the treatment and control or comparison groups is called common support. If the comparison groups do not have sufficient common support, they are not comparable; therefore, PS methods should not be used.

Even though propensity scores are the predicted probabilities of selection into a condition, the goal of PS methods is not to predict group membership, but to use propensity scores to balance treatment and control groups. The best cases for PS methods are actually those who are assigned to the treatment group, but are just as likely to be in the control group (and vice versa). Ideally, the PS distributions for both the treatment and control groups would be normal and have a mean of .5 with equal standard deviations. Under these conditions, we are most likely to replicate random assignment, as participants in one group will be very similar to those in the other group, which will allow one to obtain an unbiased treatment effect. However, we may not always see these distributions, especially when using several covariates that are strongly related to selection. Sometimes, we find that those in the treatment group have higher propensity scores than those in the control group. Therefore, in some situations, we may need to improve the common support by having more variability in the propensity scores of those in the control group compared with those in the treatment group, which may be achieved by having proportionally more participants in the control group who can be matched to those in the treatment group.

There are several methods for examining common support, such as the following: (a) making a visual inspection of the PS distributions, (b) comparing the minimum and maximum values of the propensity scores of each group, (c) using a trimming procedure (d) running an inferential test to determine if the distributions are *significantly* different from each other, or (e) estimating the mean difference of the propensity scores.

For the first method, researchers may simply graph PS distributions for the treatment and control groups and visually inspect the extent to which they overlap (Bai, 2013; Shadish, Clark, & Steiner, 2008). This can be done by comparing histograms or boxplots of the distributions of propensity scores for each group. As shown in Figure 1.1, nearly all of the propensity scores for those in the treatment group are between .03 and .5, while the propensity scores for the control group are between 0 and .8. Therefore, the area of common support (indicated by the box over the distributions) is between .03 and .5; those with propensity scores above .5 and below .03 do not have comparable matches.

**Figure 1.1** The propensity score distributions for the treatment and control groups



For the second method, one would “delete all observations whose propensity score is smaller than the minimum and larger than the maximum in the opposite group” (Caliendo & Kopeinig, 2008, p. 45). For example, if propensity scores in the treatment group range from .03 to .9 and the propensity scores in the control group range from 0 to .8, the overlapping distribution (or common support) is between .03 and .8.

The third method was used by Smith and Todd (2005), by which they identified the range of propensity scores that had a positive density within both distributions. This method not only excludes the observations in which the propensity scores do not overlap, but also drops cases with low frequencies of propensity scores in each group. For example, suppose that all overlapping propensity scores are between .03 and .8, but there are very few cases in one group or both groups with propensity scores between .5 and .8. In this approach, not only would we exclude all cases with propensity scores that were greater than .8 or less than .03, but we would also drop participants in the control group whose propensity scores were greater than .5. Likewise, if there were very few cases in the treatment group with propensity scores between .03 and .1, these cases would also be dropped (Caliendo & Kopeinig, 2008).

The fourth method consists of using an inferential statistic, such as the independent samples Kolmogorov-Smirnov test, to determine whether or not there is a significant difference between the distributions of the propensity scores for the treatment and control groups (Diamond & Sekhon, 2013). A significant difference between the two distributions would indicate poor common support. However, this method is not

recommended for the same reason that *covariate* balance should not be assessed with inferential tests: because “balance is a characteristic of the observed sample, not some hypothetical population” (Ho, Imai, King, & Stuart, 2007, p. 221). In the fifth method, researchers compute the standardized difference score ( $d = (M_T - M_C)/s_p$ ) to compare the means of the propensity scores for the treatment ( $M_T$ ) and control groups ( $M_C$ ). A small difference score (i.e.,  $d < .5$ ) indicates good common support.

Unfortunately, what constitutes *sufficient* common support is still not clear, since not all of these methods provide a clear criterion. The visual inspection using graphs and the minima and maxima comparison may provide clear criteria for cases that share common characteristics, but we do not know how much of this shared support is sufficient. While researchers have offered some guidelines, their standards are not universally recognized. For example, Bai (2015) found that selection bias is most likely to be reduced with PS matching if at least 75% of the propensity scores overlap in each distribution. If using the method of comparing the standard mean differences, Rubin (2001) recommends that the standardized mean difference between the group distributions is less than .5.

However, these general guidelines may not be sufficient when considering that the specific method of determining common support and (more importantly) of how to handle common support depends on the distributions of the data and the particular matching methods used to adjust the treatment effects. For example, if distributions are skewed or have many outliers, the inferential test or trimming procedure may assess common support better than the minima and maxima comparison or the standardized mean difference. Also, the specific method of matching will address how the degree of common support is managed. Caliper matching (see Chapter 3) uses cases with the best common support (or closest PS matches), while stratification is more lenient in its requirements for common support by allowing more flexibility in the acceptable matches. It is important to understand that how common support (or more importantly, how a lack of common support) is handled influences the validity of the results of the estimated treatment effect when using PS methods. Regardless of how common support is assessed, the defined region of common support determines which cases remain in the analyses. For instance, those cases with propensity scores outside of the range of common support may or may not be included in the final outcome analyses, depending on the specific PS method selected. If PS matching with a caliper is used, the cases with propensity scores outside of the range of common support are usually excluded in the final sample selected to estimate the treatment effect. While this restriction of cases improves the ability to match comparable cases and presumably improve internal validity, it also poses potential problems to

external and statistical conclusion validity. First, it may limit our ability to generalize the study results to the population. That is, if the cases that we dropped (i.e., those who are very likely to be selected for treatment) were systematically different from those who remained in the analysis (i.e., those who were just as likely to be selected into the treatment group as the control group), the sample selected may no longer represent its original population. Second, dropping cases will reduce the sample size, which may affect statistical power. Excluding only a few cases from a dataset with a large sample size is not problematic, but dropping half the cases from an already small sample may underpower the analysis for the treatment effect. Type II errors are just as misleading as selection bias. Therefore, if the common support is not sufficient, PS methods should not be used.

#### 1.4 Summary of the Chapter

PS methods can be effective in reducing selection bias in observational data and increasing the validity of statistical causal inference when used appropriately. More specifically, they can (a) control for multiple covariates using one composite score, (b) balance the influences from covariates on causal effect estimation when used as weights or covariate adjustments, and (c) create balanced groups that mimic those in true experimental designs. Under many conditions, PS methods are preferable to other methods used to reduce selection bias. However, it is important to note that, for propensity scores to be most effective, the conditions and assumptions that were discussed previously in this chapter must be met. The checklist below is provided to help you determine whether or not a PS method is suitable for your observational study. Assuming that it is, the next steps are to learn how to estimate and apply propensity scores. In the following chapters, we will focus on the practical applications of PS methods with an empirical example throughout the book to illustrate how to use PS methods.

#### Checklist for Using Propensity Score Methods

- You plan to examine the causal relationship between a treatment and an outcome.
- You are not certain how participants were assigned to treatment groups.
- You are familiar with theoretical or empirical evidence for why participants might choose (or be assigned to) treatment groups.

*(Continued)*

(Continued)

- You have access to several measured covariates that are related to the treatment condition and the outcome variable(s).
- The set of available covariates will include nearly all confounding factors that impact causal variables and outcomes.
- There is sufficient overlap in the PS distributions between the treatment and control groups.
- There is very little missing data within each covariate.
- Measures of the covariates are valid and reliable.

### Study Questions for Chapter 1

1. What is a group selection bias?
2. What is a propensity score?
3. When should researchers use PS methods instead of other methods to control for selection bias?
4. How do PS methods control for selection bias?
5. When might PS methods not sufficiently reduce bias?