CHAPTER SIX

COMPARISON GROUP DESIGNS

Gary T. Henry

Evaluators frequently employ comparison group designs to assess the impact of programs or policies on their intended outcomes. Comparison group designs represent alternatives to randomized experiments when the goal of the evaluation is to provide a quantitative estimate of the causal effects of a program. (Throughout, I use the term *randomized experiment* to mean a study in which target population members are randomly assigned to program participation [treatment] or a control condition [no treatment]. Such studies are also known as *randomized controlled trials* or *random assignment studies*; see Chapter Seven for more information.) The purpose of most impact evaluations is to isolate the effects of a program to help officials and citizens decide whether the program should be continued, improved, or expanded to other areas. To conduct an impact evaluation, evaluators must provide an answer to the question: What would have happened in the absence of the program? In comparison group designs, the outcomes of the target population served by the program are compared to those of another group who represent what would have occurred in the absence of the program. By contrasting the outcomes of these two groups, evaluators can estimate the impact of the program on its intended outcomes.

The need for a comparison group may be best explained by an example. Many policymakers, parents, and taxpayers want to know if publicly subsidized prekindergarten (pre-K) programs boost children's cognitive, social, and communication skills. These skills can be measured directly after children have participated in the pre-K program, but it is impossible to know from these

measures of the prekindergarteners' outcomes alone how much they were affected by the program. Even with an accurate measure of the children's problem-solving skills, it is impossible to isolate the effect of pre-K from the effects of other influences, including children's skills prior to entering pre-K, their interactions with family and others during the year, and other resources in the communities where they live. To obtain an estimate of the pre-K program's contribution to the children's outcomes, the skills of the children who participated in the program must be compared to those of another group. This is the essential logic behind comparison group designs. Although the logic of these designs is relatively straightforward, planning and implementing an evaluation that *accurately* isolates the effects of a program on the target population is one of the most difficult challenges that evaluators undertake.

Some readers may already be wondering about the definition of *comparison group design*. Fundamentally, comparison group designs are approaches to assessing causal effects that measure an important program outcome and estimate impact by comparing the difference between treated and untreated groups when random assignment to the groups has *not* been used. Thus the term *comparison group* distinguishes an untreated but not randomly assigned group from a randomly assigned *control group* used in randomized experiments. Both control and comparison groups are used to estimate what would have happened in the absence of a program, but group members are selected differently.

This chapter describes a variety of comparison group designs that evaluators have frequently used to assess program impacts. The tasks involved in developing and carrying out a comparison group design that accurately and convincingly estimates the effects of a program are arduous. Fortunately, in the past few years a stronger base of theory and empirical studies that examine the accuracy of alternative comparison group designs has begun to emerge. However, it is important to understand at the outset that (1) none of the comparison group designs is without its limitations; (2) some are better than others; and (3) in a world in which the most convincing estimate of program impacts is *not the top priority* for most policymakers and the public, evaluators will need to be knowledgeable and resourceful in exploring alternative comparison group designs when planning impact evaluations. Also, evaluators need to be able to explain the strengths and limitations of each design to sponsors and stakeholders so that informed choices can be made and evidence concerning program impacts can be accurately interpreted.

Comparison group designs have previously been grouped together under labels such as quasi-experimental, pre-experimental, non-experimental, or

observational studies. They also include some designs that are possible when evaluators have longitudinal data, such as administrative data that tracks program participants over time or studies that collect data on multiple cohorts that participate in a program. This chapter strips away the distinctions among these labels by referring to all designs for impact evaluations that do not involve random assignment to treatment and control as *comparison group designs.* The unifying concept is that all comparison group designs involve using the difference between treated and untreated comparison groups in order to estimate the effects of a program. The idea behind comparison group designs is simple, but the program impact estimates can be quite misleading if the conditions for implementing them are not present or they are carried out incorrectly. Only in a limited number of situations, when circumstances are favorable and the comparison group designs are implemented in specific ways, have studies shown that the program impact estimates are sufficiently accurate for potential issues with the design to be ignored (Bifulco, 2012; Cook, Shadish, and Wong, 2008; Glazerman, Levy, and Myer, 2003; Shadish, Clark, and Steiner 2008). However, in an imperfect world, evaluators must sometimes choose between better and worse designs rather than the best design or no evaluation. The goal of this chapter is to help practicing evaluators improve their designs as much as circumstances and resources permit and be able to carefully state the limitations on the findings for any comparison group study that they carry out.

In the next section of the chapter, I describe the ideal situation for assessing program impacts. Surprisingly to some, this ideal is not a randomized experiment. In the view of many evaluators, statisticians, and social scientists, randomized experiments are the closest approximation to the ideal that is possible. However, randomized experiments are not the ideal, as we shall see, and they require that certain assumptions be made to consider their estimates truly accurate or unbiased. Randomized experiments are usually the best potential design because they require the fewest assumptions to be made in order to consider the estimates that they produce accurate and convincing. Under certain circumstances, however, comparison group designs can produce accurate estimates of program impact; indeed, at least one of the designs is considered by many to be as compelling as randomized experiments. Having acknowledged that comparison group designs can produce accurate effect estimates, it is important to also say that these designs impose significant technical conditions and data requirements that must be met in order to produce accurate and convincing estimates of program impact.

Introduction to Causal Theory for Impact Evaluation

Impact evaluations seek to provide quantitative estimates of the causal effects of programs or policies. The potential outcomes framework provides a new way to think about this task. The ideal way to calculate the causal impact of a program is to compute the difference between the outcome for every member of the target population after participating in a program *and* the outcome for the same subjects after not participating in the program. Applying this idea to the earlier example, an impact evaluator would like to be able to observe the outcomes for a specific four-year-old after attending prekindergarten *and* after not attending prekindergarten. Of course, each child either attends or does not attend a pre-K program when he or she is four years old. So in reality we can never directly observe or measure the treatment effect, a situation that Paul Holland (1986) refers to as the "fundamental problem of causal inference." Because we need to be able to acquire actual evidence about the impacts of policies and programs, evaluation designs that approximate the ideal have been developed and tested.

The most ideal alternative to this impossible situation for generating an unbiased effect estimate is a study that would involve random sampling and random assignment (Imai, King, and Stuart, 2009; Murnane and Willett, 2012). In such a study, a random sample of the target population for a program would be selected, then each of the randomly selected sample members would be randomly assigned to treatment or control. Why would this be considered an ideal? Random sampling ensures that the sample is an accurate statistical model of the population or has external validity (Henry, 1990). Random assignment of individuals to the treated groups or untreated control groups ensures that all the other factors that influence the outcomes are unrelated to whether the individual is treated or untreated. In other words, random assignment to one or the other group can be expected to balance all of the influences on the observed outcomes after program participation *other than the program.* When individuals are randomly assigned to treatment, differences between the treatment group and control group means on outcomes can be reasonably attributed to the effects of the program, when chance variation can be dismissed as an explanation, and the sample has been selected at random from the target population, the effects can be inferred to the target population. Comparison group designs are usually implemented in situations when random assignment is not possible, and therefore we will focus on those situations.

In the world of practical program evaluations, random assignment is always difficult and frequently impossible. A principal advantage of random assignment in impact evaluations is that the means by which individuals enter the treated or untreated groups is free of selection bias or, in other words, assignment to treatment is not related to the outcomes of the program participants. For example, if a fair lottery is used to assign eligible children to Pre-K when the demand for slots exceeds availability, then comparing outcomes of the children who received a slot and those who did not is free of selection bias. If, however, the spaces are allocated on a first-come first-served basis, parents who are more motivated for their children to succeed in school may be more successful in enrolling their children. After the program, any differences in outcomes measured for the four-year-olds who participated and the four-year-olds who did not participate (who were untreated) could be due in part to parental motivation as well as in part to the program. It is wise for evaluators to assume that selection bias does operate when designing impact evaluations, even if some argue that it is nonexistent or of minimal consequence. Comparison group designs differ in the extent to which they can minimize this type of bias.

Randomized experiments stand out as the preferred method for eliminating selection bias from impact evaluations. However, other sources of bias can and often do occur with randomized experiments, such as non-random selection from the target population when those who volunteer for participation are different from the target population that is of interest to evaluation stakeholders, when participants leave the study, or when samples are too small to rule out chance as an explanation for observed differences in outcomes. For these reasons and for a number of practical concerns—such as the length of time it takes to design, implement, and obtain findings from a randomized experiment, the costs of randomized experiments, and ethical issues—comparison group designs are useful and powerful approaches for evaluators to consider when designing impact evaluations. Often, comparison group designs are the only practical means available for evaluators to provide evidence about program impact within time and resource constraints imposed by evaluation sponsors.

Comparison Group Designs

Numerous comparison group designs are available for impact evaluations. These designs differ in their data requirements, the circumstances in which they can be applied, and the extent to which they are likely to reduce bias. Although theory (Holland, 1986; Rosenbaum and Rubin, 1983) and careful

application of logic (Shadish, Cook, and Campbell, 2002) provide guidance about the relative strengths of alternative comparison group designs, neither theory nor logic can predict the extent of bias that any particular impact evaluation design will produce. Our understanding about the extent of selection bias in any specific evaluation is largely based on judgments about the plausibility of alternative explanations for the results and on empirical studies that compare program impact estimates of randomized experiments and comparison group designs.

The next eight sections of this chapter review alternative comparison group designs, beginning with the most basic design, which is labeled the *naïve* design. Even though this design is the least likely to produce estimates that are accurate and convincing, it illustrates the most elementary approach to comparison group design. These eight sections end with a review of a very rigorous and sophisticated design—*regression discontinuity*—that has strong theory supporting the accuracy of its causal inferences and mounting empirical evidence about its accuracy. In between these two extremes are additional comparison group designs that are described roughly in order of their likelihood of producing more and more accurate impact estimates, but also in order of increasing complexity and greater requirements for data and technical proficiency. I say roughly in order of each design's ability to yield accurate estimates of program impact because this accuracy depends on the circumstances in which the design is used, the data that are used for the evaluation, and how well the design is implemented. A poorly implemented randomized experiment can produce flawed findings, and a poorly implemented comparison group design can as well. None of the designs are in any sense guaranteed to produce accurate impact estimates, but in moving us in the direction of providing evidence about program effects, they represent efforts to provide the best available evidence about program effects when randomized experiments are infeasible, unethical, or not available in the time frame in which decisions need be made. Finally, it is important to note that elements of these designs can be combined to enhance the accuracy of program impact estimates, although describing the possible combinations is beyond the scope of this chapter.

1. Naïve Design

The most basic means of estimating program impacts is the comparison of the statistical means for *naturally occurring* treated and untreated groups. Usually, this design involves comparing one or more outcome measures for a group that participated in or had access to treatment with those for another group that did not have such treatment or access to it. The outcome measures may

be available from administrative data, such as student test scores or patient lengths of stay in the hospital, or from surveys that include both treated and untreated individuals. To continue with the pre-K example, we could compare children who participated in pre-K and those who did not on teachers' ratings of their school readiness when they begin kindergarten. An estimate of impact is the difference between the group means:

$$\delta = \overline{Y}_t - \overline{Y}_u$$

In this case, δ is the treatment effect or program impact (for example, the pre-K program impact estimate),

$$\overline{Y}_t$$

is the mean score on the readiness assessment for the treated (for example, prekindergarteners), and

$$\overline{Y}_u$$

is the mean score for the untreated (for example, children who did not attend pre-K). Naïve comparisons can also be expressed in a simple regression format (this chapter also uses this format for depicting other designs, and therefore it is shown here for clarity):

$$Y_i = \alpha + \delta Z_i + \varepsilon_i$$

where $Y_i$is the outcome (for example, the readiness score for each kindergartener), α is the average score for the untreated (for example, kindergarteners who did not participate in pre-K), δ is a regression coefficient that represents the difference between the treated and untreated group means and is identical to the δ in the previous equation,

$$Z_i$$

is an indicator variable that is coded 1 for treated and 0 for untreated individuals, and

$$\varepsilon_i$$

is each individual's deviation from the group mean, also described as the *error term.*

This design for producing program impact estimates when the treated and untreated groups have occurred naturally is considered naïve because these two groups could differ for reasons other than participation in the program. The treated group could have been in greater need of the program and started out far behind the untreated group in skills, family resources, or motivation. In some situations the reverse might be expected. For example, families or individuals who perceive greater benefits from program participation, like pre-K, after-school, or job training programs, may have been the most likely to enroll.

Frequently, in addition to individuals making a choice to participate (self-selection), program operators choose individuals to participate based on such loose criteria as the operator's perceptions about need or the extent to which an individual will benefit from the program. In situations where participants or program operators exercise judgments in selecting program participants, selection bias is likely to render the simple comparisons highly inaccurate. In any case, after program participation has occurred, it is impossible to discern conclusively that selection bias did not occur. Selection bias is the most pernicious form of bias that plagues evaluators' attempts to accurately estimate program impacts. Also, it is often difficult to explain to evaluation sponsors and stakeholders why this simple design is unlikely to be accurate.

We can gain a better understanding of the reasons why bias is likely to be a problem by recourse to the *potential outcomes framework*: for selection bias to be reduced to the point at which it can be ignored, *the potential outcomes if neither group was exposed to the program* must be similar for the two groups. Evaluators must ask the following question and be able to answer it in the affirmative for the naïve design to be a reasonable approach for an impact evaluation: Do we expect that if the program had not been offered, the outcomes of the treated and untreated comparison groups would have been roughly the same? In other words, evaluators must be able to assume that if no treatment had been offered to anyone in either group, the mean outcomes of the treated and untreated groups would have been the same *and* that assumption must be both valid and plausible to the consumers of the evaluation.

In most cases, good and valid reasons exist to doubt that the outcomes would have been roughly the same. Moreover, evaluators and evaluation sponsors and stakeholders cannot be sure that the outcomes would be roughly the same for both groups had the program not been provided. Thus, in the great majority of cases, the assumption of equivalent outcomes for the treated and untreated groups in the absence of the program is likely to be implausible and

naïve. Therefore, the naïve design is a very questionable choice for impact evaluations in most cases.

2. Basic Value-Added Design: Regression Adjusted for a Preprogram Measure

As mentioned previously, one of the difficulties of the naïve design is that evaluators cannot rule out the possibility that differences between the treated and untreated groups prior to participating in the program could cause differences in their outcomes after the program has been offered to the treated group. When the same measure that is used for the outcome has been observed for both the treated and untreated group *prior to program participation*, the preprogram measure can be used to adjust the impact estimates for this preexisting difference. Adding this preprogram measure to the regression equation adjusts the program impact estimates for preexisting differences between treated and untreated on an outcome of interest for the evaluation. Here is a regression approach to estimating program impact (δ) using a preprogram measure of the outcome:

$$Y_{it} = \alpha + \delta Z_i + \beta_1 Y_{it-1} + \varepsilon_i$$

As in the previous equations, δ represents the program impact estimate, Z_i is coded 1 for the treated group members and 0 for the untreated, Y_{it} is the measure of the outcome at time t, which occurs after treatment, Y_{it-1} is the measure of the outcome at time $t-1$, which occurs before the program, and α is the constant or intercept term. The preprogram measure is the best measure of the individuals' differences prior to the program and therefore adjusts the program impact for the initial difference in the outcome variable between the treated and untreated groups within limitations of measurement error and assuming that the values of the preprogram measures for the two groups substantially overlap.

This comparison group design of program impact adjusting for a preprogram measure of the outcome variable is commonly referred to as a *value-added design*, because the program impact estimate represents the amount of gain since program participation began or, more precisely, since the preprogram measure was observed. Fortunately, this type of longitudinal design is now possible more frequently due to administrative data sets that record the status of individuals eligible for a program at regular intervals, such as the test scores measures that are now available for most students in all states. In addition, much has been done to assess the accuracy of value-added approaches and

the threats to the accuracy of the estimates that are generated. Common problems with value-added designs include regression to the mean (high and low scores at preprogram move toward the mean after the program), differences in program effects for those with either high or low preprogram scores, and omission of other covariates that are related to both program participation and the outcome. The latter problem, which is sometimes called *omitted variable bias*, is a persistent problem with comparison group designs. Omitted variable bias, which refers to the bias of program impact estimates due to omitting other variables that are related to the outcome and program participation, will be confronted but not entirely removed in the next design, the regression-adjusted covariate design.

3. Regression-Adjusted Covariate Design

A frequent first impulse to improve program impact estimates over the naive estimates is to use multiple regression to adjust the impact estimates by adding covariates (control variables) that account for some of the other variables that influence outcomes. To the extent that the treated and the untreated groups differ on covariates systematically associated with the outcome, the estimates of program impact can be adjusted for the imbalances between the groups. This chapter calls this design *regression adjusted for covariates*, and it can be represented as follows:

$$Y_i = \alpha + \delta Z_i + \beta_1 X_{1i} + \cdots + \beta_c X_{ci} + \varepsilon_i$$

where β_1 is the regression coefficient for the first of c covariates, X_{1i} is the value of the first of c covariates for individual i, X_{ci} is the value for the last of the c covariates, and δ is the adjusted coefficient that is used to estimate the program impact. When samples are large enough, many covariates can be added to the multiple regression equation. The choice of covariates should be guided by theory that delineates the various factors that influence the outcomes of interest. Where possible, it is advantageous to use empirically validated theory that prior research has shown to be systematically associated with the outcomes in the target population for the program being evaluated. Where the covariates are better at explaining the variation in the outcome and differ between treated and untreated groups, the bias reduction is likely to be significant.

Studies (Glazerman, Levy, and Myer, 2003; Shadish, Clark, and Steiner, 2008) find that this type of design does substantially reduce bias; however, their findings also indicate that substantial bias can still remain, depending on the quality of the covariates. For example, in an evaluation of a state pre-K

program, Henry and Rickman (2009) report that 51 percent of the mothers of Head Start participants received food stamps, but only 18 percent of the mothers of children enrolled in Pre-K received food stamps. An adjustment for the difference could be attempted by adding food stamp receipt as a covariate in an equation that includes the indicator for participation in Pre-K. In addition to food stamp participation, additional covariates, variables on which the two groups could be expected to differ and which are expected to affect the outcome, could be added to adjust the estimates further. Larger reductions in bias may be achieved when the covariates include all variables that have affected selection into treatment and the outcomes that are being assessed. However, the extent of bias remaining cannot be directly tested in most evaluations, and for that reason, covariate-adjusted estimates of program impact are frequently viewed with skepticism. A large number of covariates based on a well-conceived theory about the factors influencing the outcome variables will improve this design.

4. Value-Added Design Adjusted for Additional Covariates

When a preprogram measure of the variable used as a post-program outcome is combined with additional covariates, the design can be referred to as a *value-added design adjusted for additional covariates.* For example, in the evaluation of a state Pre-K program, Henry, Gordon, Henderson, and Ponder (2003) measured children's cognitive and communication skills on four different assessments in fall 2001, when the children began Pre-K, Head Start, or some other preschool program. Because these same measures were used as the outcome variables after the Pre-K program, they could be added along with other baseline covariates expected to explain a great deal of the variation in children's developmental outcomes into the regression analysis, which can be represented as follows:

$$Y_{it} = \alpha + \delta Z_i + \beta_1 Y_{it-1} + \beta_2 X_{1i} + \cdots + \beta_{c+2} X_{c+1i} + \varepsilon_i$$

where the covariates include Y_{it-1} to represent the preprogram measure of the outcome, as it did in the value-added design above. The preprogram outcome measure is also known as the *lagged value* of the outcome variable. The preprogram score on the same measure as the outcome provides an adjustment for differences in the children in the treated and untreated comparison groups on that measure before the participation in the program began for the treated group. In addition, it can improve the precision of the estimates of treatment effects. Including the preprogram score in the equation changes

the interpretation of the coefficients, $\beta_2 - \beta_{c+2}$, to the effect of the covariate since the time the preprogram score was measured. Some recent research has indicated that estimates are likely to be improved if all the preprogram measures, such as measures of skills for four-year-olds, are included in the regression. Doing this appears to mitigate a part of the measurement error that can occur in any single measure of the children's skills and abilities. To continue with this example, preprogram measures can also be used to adjust for other outcome measures, such as teachers' ratings of school readiness, but the coefficients are not interpreted as value added.

When a preprogram measure of the outcome variable is available, it is possible to develop a special type of covariate-adjusted, value-added program impact estimate that substantially reduces bias (Glazerman, Levy, and Myer, 2003). The process for this special adjustment begins by assembling a rich set of covariates and using the preprogram score on the outcome variable as a temporary substitute for the post-treatment outcome variable in a regression analysis. First, the preprogram score is substituted as the dependent variable, as shown here:

$$Y_{it-1} = \alpha + \delta Z_i + \beta_1 X_{1i} + \cdots + \beta_c X_{ci} + \varepsilon_i$$

The object of the analysis is to find a set of covariates that render the coefficient on the indicator for treatment (δ) statistically insignificant (that is, not statistically different from 0). This indicates that the covariates included in the equation make the two groups statistically similar prior to program participation. Then the program impact is estimated using the equation for the value-added design adjusted for covariates presented earlier in this section, with the complete complement of covariates included in the specification in which the treated and untreated groups were found to be statistically more similar.

5. Interrupted Time-Series Designs

Interrupted time-series designs are longitudinal designs applicable in situations when new programs are started or major reforms occur and data are available for before and after the change. These situations are often referred to as *natural experiments*, but they compare groups that cannot be assumed to be equivalent, as groups in randomized experiments can be. Interrupted time-series designs include several observations of an outcome that is expected to be affected by the reform or new program *prior to the change* and several observations of the same measure *after the change*. In the most basic form of interrupted time-series designs, the treated group before the reform is compared to itself

after the reform. However, this design is considered relatively weak because other changes occurring at the same time as the reform might produce a different pattern in the outcome variable after the interruption.

A much stronger interrupted time-series design involves comparing post-program change from the pattern of preprogram observations in the treated group to the change for an untreated group over the same time. This design is frequently assessed for an average increase in the outcome during the post-program period in comparison to the change in the untreated group's outcome, which can be done using the following formula:

$$Y_{it} = \alpha + \beta_1 Z_i + \beta_2 T + \delta Z_i T + \beta_3 X_{ci} + \cdots + \beta_{c+3} X_{c+1i} + \varepsilon_i$$

where the estimate of the effect is the coefficient δ on an interaction term that designates observations of the treated group after the reform or change, Z_i is coded 1 for the treated group and 0 for the untreated group, and T_t is coded 1 for the period after the interruption and 0 for the pre-interruption period. The program impact coefficient is interpreted as the average increase (or decrease, if negative) in the outcome for the treated group in the post-interruption period less the average difference in the outcome for the untreated group in the post-interruption period.

This interrupted time-series design is considered relatively strong. The principal issue with the design is that the treated group could have been selected based on criteria that make them react to the reform or program change differently from the untreated group. For example, this design has been used to assess the impact of various types of state policy reforms on the intended outcomes of the policy. But if a state adopts a reform of a particular policy in a period when pre-interruption outcomes are particularly low, then factors other than the policy may be contributing to any increase that is found.

Impact evaluations are currently making use of this design more frequently. For example, Henry and Gordon (2003) assessed the effects of a public information campaign that announced action days during which members of the Atlanta public were asked to reduce car travel due to predictions that ozone levels would exceed U.S. Environmental Protection Agency regulations. Upcoming action days were communicated to the public on the day prior to the predicted violation through electronic highway signs and in weather forecasts in the newspapers and in radio and television broadcasts. Thirty-five action days occurred intermittently during the 153-day study period, and so driving behavior on the action days was contrasted with driving behavior on the other days. Rather than a single switch or interruption, this design used multiple switches or interruptions to estimate the program effects. Henry and

Gordon found that overall, individuals reduced their average miles driven on action days by 5.5 miles, and individuals whose employers had joined in the local clean air campaign reduced their driving even more.

6. Fixed-Effect Designs for Longitudinal Evaluations

With longitudinal data sets becoming more commonly available for studies such as education and labor market studies and for panel studies such as the Panel Study of Income Dynamics and the Early Childhood Longitudinal Studies, a study design known as a *fixed-effect design* has gained popularity. In these designs each individual effectively serves as his or her own comparison. The treatment effects are estimated based on comparing outcomes for an individual during periods when he or she received a particular treatment with outcomes during periods when he or she received an alternative treatment or no treatment.

$$Y_{it} - \overline{Y}_i = \alpha + \delta\left(Z_{it} - \overline{Z}_i\right) + \beta_c\left(X_{cit} - \overline{X}_{ci}\right) + \left(\varepsilon_{it} - \overline{\varepsilon}_i\right)$$

The individual fixed effects remove all of the non-time varying characteristics of the individual, which essentially allows for each individual to serve as his or her own control. An evaluation of the Teach For America (TFA) program, by Xu, Hannaway, and Taylor (2008) illustrates individual fixed effects. The objective of the evaluation was to obtain the effects of TFA by estimating the differences in student achievement when students were taught by a TFA teacher and when they were taught by a traditionally prepared teacher. The evaluation used data from the twenty-three school districts in North Carolina where TFA teachers were employed between the 2000–2001 and 2006–2007 school years. Treatment effects were the aggregated differences between end-of-course achievement test scores for courses taught by TFA teachers and end-of-course test scores for courses taught by traditionally prepared teachers for each student who had experienced both types of teachers. The student fixed effects essentially eliminated the influence of any student characteristic that did not change during the study period. Xu, Hannaway, and Taylor (2008) found that students in courses taught by TFA teachers averaged scores about one point higher on their tests (0.10 standard deviation units) than they did when these students were taught by other teachers. Although fixed effects can eliminate an important source of bias, only study subjects who have experienced treatment on some occasions and not experienced it on others can be included in impact estimates. In the example, students who did not experience

both types of teachers were excluded from the effect estimate, which may limit the generalizability of the effect estimate.

7. Matching Designs

Designs in which members of an untreated group are matched to each member of a treated group are referred to as *matching designs*. Matching members of the comparison group to members of the treatment group using preprogram covariates is intended to make the two groups as similar as possible prior to the treatment. These designs have a venerable but checkered tradition in evaluation. There are instances in which matching did not produce results that most careful observers would regard as unbiased. One such case was the now infamous Westinghouse study of Head Start (Cicarelli, 1971). In this evaluation, children who had participated in Head Start were matched with members of their kindergarten classes who had not participated, and differences in outcomes between these two groups were used to estimate the effects of Head Start. However, when the data were reanalyzed using several different techniques, the evaluators concluded that substantial bias was likely (Datta, 1982). The issue for matching designs is that the groups can be made very similar on the covariates that have been measured and used in the matching, but unobserved differences in the treated and untreated groups may remain, and these differences may bias the impact estimates.

One strategy for reducing the differences between a treated group and an untreated comparison group is to increase the number of covariates that are used to match them. However, as more covariates are added it becomes difficult to find exact matches for every member of the treated group in the pool of untreated group members. The result is inexact matches or incomplete matches, either of which could affect the accuracy of the impact estimates. Recent developments in matching designs indicate that an approach referred to as *propensity score matching* may substantially reduce bias in effect estimates (Cook, Shadish, and Wong, 2008; Diaz and Handa, 2005; Glazerman, Levy, and Myer, 2003; Stuart 2010). Paul Rosenbaum and Donald Rubin developed the propensity score matching approach, and they have established some very desirable properties for the procedure (see Rosenbaum and Rubin, 1983; Rosenbaum, 2002; Rubin, 2001).

Several different types of matching designs have become available since compelling theoretical properties for propensity score matching (PSM) were developed by Rosenbaum and Rubin (1983). PSM replaces matching on several covariates with matching on a single propensity score that is an estimate of the likelihood that each member of the study sample could be in the

treated group. PSM makes assignment to the treated and comparison groups independent of the covariates that are used to estimate the propensity of an individual to be in the treatment group. PSM has both intuitive and technical advantages over using regression-based approaches by themselves. These advantages include creating a comparison group that is more similar to the treatment group and accounting for nonlinearities between covariates and the outcome (Stuart and Green, 2008). However, the theory supporting PSM does not address the extent of the bias that could be caused by imbalances between treatment and comparison groups that result from variables not included in estimating the propensity score. Therefore, the amount of bias that remains after the matching is uncertain. If assignment to the two groups is not independent across the unobserved covariates, the treatment effect estimates using PSM could be biased, due to these differences between the treated and comparison groups that are uncontrolled by the matching.

Propensity score matching has been frequently used in recent years to obtain more accurate estimates of program impact. In a recent study, Henry, Gordon, and Rickman (2006) matched a group of children who participated in a state Pre-K program with a group of Head Start recipients to investigate the differences in developmental outcomes attributable to participating in Head Start when compared with an alternative Pre-K program. They found that the children who attended the state Pre-K program scored higher on direct assessments of language and cognitive skills at the beginning of kindergarten than the children who participated in Head Start.

When PSM procedures are followed, within-study comparisons indicate significant bias reduction (Cook, Shadish, and Wong, 2008; Diaz and Handa, 2005; Glazerman, Levy, and Myer, 2003; Stuart 2010). However, some researchers find the amount of bias that remains is too large to ignore (Glazerman, Levy, and Myer, 2003; Wilde and Hollister, 2007). There are a variety of ways to obtain estimates of the treatment effect using the matched sample. A common approach, which Glazerman, Levy, and Myer (2003) find to reduce bias, simply adds the propensity score (P_i) to the regression equation used for the estimation of the treatment effect, as illustrated in this equation, which assumes that a lagged outcome measure is available:

$$Y_i = \alpha + \delta Z_i + \beta_1 Y_{it-1} + \beta_2 P_i + \beta_3 X_{1i} + \cdots + \beta_{c+3} X_{c+1i} + \varepsilon_i$$

The accuracy of program impact estimates using propensity score matching remains an empirical question because the influence of unobserved variables depends on the quality of the covariates used in the estimation of the

propensity score and in the previous equation. Having a large group of covariates that are related to the outcome and selection into the treated group has been shown by empirical research to be important for the accuracy of the program impact estimates.

8. Regression Discontinuity Designs

Regression discontinuity (RD) designs lay the strongest claim of all of the comparison group designs to producing unbiased estimates of program impacts. These designs rely on a quantitative index to assign subjects to treatment or to an alternative condition, such as business as usual or no treatment. Including the quantitative index in the equation used to estimate the difference between the treated and comparison groups can effectively eliminate selection bias in the program impact estimates (Imbens and Lemieux, 2007; also see Chapter Seven in this volume). Frequently, the quantitative variable that assigns subjects to treatment measures an individual's need for the program or the subject's risk of adverse social consequences. The variable does not have to perfectly measure the need, but it must be strictly used to assign all cases on one side of a cutoff value to the treated group, and the treatment must be withheld from all cases on the other side of the cutoff.

Cook, Shadish, and Wong (2008) provide evidence that the differences in estimates from RD designs and randomized experiments are negligible. RD design produces impact estimates that are unbiased by selection into treatment by including the variable that indicates whether a subject was assigned to the treated group or comparison group and the quantitative index used to assign subjects to the two groups along with higher-order terms of this variable in a multiple regression (Imbens and Lemieux, 2007, p. 10). Including variables in the estimation equation that perfectly model the relationship between the assignment index and the outcome variable breaks up any possible correlation between the treatment indicator variable and the individual error term. Including this term along with exponential functions of this term as needed removes selection bias from the estimate of the treatment effect.

Here is a typical equation for estimating the treatment effect (δ):

$$Y_i = \alpha + \delta Z_i + f(I_i - I_c) + \beta_1 Z_i * f(I_i - I_c)_i + \beta_2 X_{1i} + \cdots + \beta_{c+2} X_{c+1i} + \varepsilon_i$$

where the novel term, $f(I_i - I_c)$, can be broken down as follows: I_c is the value of the assignment index at the cutoff, I_i is the value of the index for each individual, which makes the term $I_i - I_c$ the distance of each individual from the cutoff. The entire term refers to a polynomial function of the distance, which

is operationalized by including exponential functions such as the squared and cubed terms of the distance from the cutoff on the assignment variable as additional variables until the higher order terms are no longer significant. Basically, the terms associated with the assignment index I will "soak up" the selection bias that may be attributable to omitted variables, so that even though the *other* coefficients may be biased, the program impact estimate will be correspondingly purged of selection bias. The covariates are included to eliminate small sample biases and increase precision without altering the benefits of the RD design (Imbens and Lemieux, 2007, p. 11).

In an evaluation of the impact of Oklahoma's state pre-K program, Gormley, Gayer, Phillips, and Dawson (2005) used a novel RD design to estimate the effects of the program in Tulsa. The evaluation team used the children's ages as the assignment variable and the eligibility dates as the cutoff on the assignment variable. The children's outcomes on direct assessments were measured in the fall after the older children had attended pre-K and before pre-K began for the younger children. The evaluators compared the test scores of children with ages slightly above the cutoff to the test scores of children slightly too young to have been eligible for pre-K the prior year. Significant differences in children's knowledge of letters and words, spelling skills, and problem-solving skills were found, indicating that the Tulsa pre-K program was effective in increasing these outcomes. For a more complete discussion of this design see Lipsey, Weiland, Yoshikawa, Wilson, and Hofer (2014).

RD designs have also been implemented when a few cases close to the cutoff have ended up in a group other than the one to which they should have been assigned. This type of RD design is referred to as a *fuzzy* discontinuity design, as opposed to a *sharp* discontinuity design. In addition, RD designs have been implemented in situations in which settings or clusters of individuals, such as school districts, are assigned to treatment using a quantitative index of need. Currently, the interest in RD designs is high and their use is enjoying a renaissance. A well-designed RD study effectively eliminates selection bias, but researchers must take care to assess other sources of bias. A primary validity concern for RD designs is the generalizability of the intervention effect estimate. RD designs estimate the effect of the intervention at the cutoff, and this estimate can be different from the average treatment effect estimate obtained in random assignment studies. Other sources of bias are "hidden" treatments for either group that occur precisely at the cutoff for assignment. Even with these potential biases, RD designs are generally considered second only to randomized experiments in terms of producing accurate estimates of program effects.

Conclusion

A common use of comparison group designs in evaluation is to estimate the impact of programs by calculating the difference between the outcomes for those exposed to a treatment and the outcomes for those in an untreated comparison group. When random assignment is impossible or impractical, a comparison group design can be used to contrast the outcomes of a group that participated in a program with the outcomes of an untreated comparison group in an attempt to quantify the program's causal effects from the comparison of the two group means. This chapter showed how stronger comparison group designs can be used to improve on naïve comparisons by making adjustments that reduce bias in the estimates of program impacts.

Value-added designs, interrupted time-series designs, fixed-effect designs, matching designs, and regression discontinuity (RD) designs appear to have the capacity to significantly reduce the bias likely to occur in naïve program impact estimates. Regression discontinuity designs have the most compelling underlying theory for causal inference. As the developments in propensity score matching and RD designs indicate, this is a dynamic area for research and development. As richer data sources continue to become more available, evaluators will have increasing opportunities to implement stronger comparison group designs. In addition, these designs are frequently being combined – such as in the use of matching and interrupted time series (Henry, Smith, Kershaw, and Zulli, 2013) or longitudinal data and regression discontinuity (Wing and Cook, 2013) to further reduce bias in causal effect estimates.

References

Bifulco, Robert. "Can Nonexperimental Estimates Replicate Estimates Based on Random Assignment in Evaluations of School Choice? A Within-Study Comparison." *Journal of Policy Analysis and Management*, 2012, *31*(3), 729–751.

Cicarelli, V. "The Impact of Head Start: Executive Summary." In F. Caro (ed.), *Readings in Evaluation Research.* New York: Russell Sage Foundation, 1971.

Cook, T. D., Shadish, W. R., and Wong, V. C. "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management*, 2008, *27*(4), 724–750.

Datta, L.-E. "A Tale of Two Studies: The Westinghouse-Ohio Evaluation of Project Head Start and the Consortium for Longitudinal Studies Report." *Studies in Educational Evaluation*, 1982, *8*, 271–280.

Diaz, J. J., and Handa, S. "An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from Mexico's PROGRESA Program." *Journal of Human Resources*, 2005, *41*(2), 319–345.

Glazerman, S., Levy, D. M., and Myer, D. "Nonexperimental Versus Experimental Estimates of Earnings Impacts." *Annals of the American Academy of Political and Social Science*, 2003, *589*, 63–93.

Gormley, W. T. Jr., Gayer, T., Phillips, D., and Dawson, B. "The Effects of Universal Pre-K on Cognitive Development." *Developmental Psychology*, 2005, *41*(6), 872–884.

Henry, G. T. *Practical Sampling*. Thousand Oaks, CA: Sage, 1990.

Henry, G. T., and Gordon, C. S. "Driving Less for Better Air: Impacts of a Public Information Campaign." *Journal of Policy Analysis and Management*, 2003, *22*(1), 45–63.

Henry, G. T., Gordon, C. S., Henderson, L. W., and Ponder, B. D. "Georgia Pre-K Longitudinal Study: Final Report 1996–2001." Atlanta, GA: Georgia State University, Andrew Young School of Policy Studies, 2003.

Henry, G. T., Gordon, C. S., and Rickman, D. K. "Early Education Policy Alternatives: Comparing the Quality and Outcomes of Head Start and State Pre-Kindergarten." *Educational Evaluation and Policy Analysis*, 2006, *28*, 77–99.

Henry, G. T., and Rickman, D. K. "The Evaluation of the Georgia Pre-K Program: An Example of a Scientific Evaluation of an Early Education Program." In K. Ryan and J. B. Cousins (eds.), *The Sage International Handbook of Educational Evaluation*. Thousand Oaks, CA: Sage, 2009.

Henry, Gary T., Smith, Adrienne A., Kershaw, David C., and Zulli, Rebecca A. "Formative Evaluation: Estimating Preliminary Outcomes and Testing Rival Explanations." *American Journal of Evaluation*, 2013, *34*, 465–485.

Henry, G. T., and Yi, P. "Design Matters: A Within-Study Assessment of Propensity Score Matching Designs." Paper presented at the meeting of the Association for Public Policy Analysis and Management, Washington, D.C., Nov. 7, 2009.

Holland, P. W. "Statistics and Causal Inference." *Journal of the American Statistical Association*, 1986, *81*, 945–960.

Imai, Kosuke, King, Gary, and Stuart, Elizabeth A. "Misunderstandings Between Experimentalists and Observationalists About Causal Inference." *Journal of the Royal Statistical Society*, 2008, *171*, 481–502.

Imbens, G. W., and Lemieux, T. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*, 2007.

Lipsey, M. W., Weiland, C., Yoshikawa, H., Wilson, S. J., and Hofer, K. G. "The Prekindergarten Age-Cutoff Regression-Discontinuity Design: Methodological Issues and Implications for Application." *Educational Evaluation and Policy Analysis*, 2014.

Murnane, Richard J., and Willett, John B. *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. New York: Oxford University Press, 2012.

Rosenbaum, P. R. "Covariance Adjustment in Randomized Experiments and Observational Studies (with Discussion)." *Statistical Science*, 2002, *17*, 286–327.

Rosenbaum, P. R., and Rubin, D. B. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 1983, *70*(1), 41–55.

Rubin, D. B. "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation." *Health Services & Outcomes Research Methodology*, 2001, *2*(1), 169–188.

Shadish, William R., Clark, M. H., and Steiner, Peter M. "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *Journal of the American Statistics Association*, 2008, *103*(484), 1334–1356.

Shadish, William R., Cook, T. D., and Campbell, D. T. *Experimental and Quasi-Experimental Designs.* Boston, MA: Houghton Mifflin, 2002.

Stuart, E. A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, 2010, *25*(1), 1–21.

Stuart, E. A., and Green, K. M. "Using Full Matching to Estimate Causal Effects in Non-Experimental Studies: Examining the Relationship Between Adolescent Marijuana Use and Adult Outcomes." *Developmental Psychology*, 2008, *44*(2), 395–406.

Wilde, E. T., and Hollister, R. "How Close Is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment." *Journal of Policy Analysis and Management*, 2007, *26*(3), 455–477.

Wing, C., and Cook, T. D. "Strengthening the Regression Discontinuity Design Using Additional Design Elements: A Within-Study Comparison." *Journal of Policy Analysis and Management*, 2013, *32*, 853–877.

Xu, Z., Hannaway, J., and Taylor, C. *Making a Difference? The Effects of Teach For America in High School.* Washington, DC: Urban Institute and CALDER, 2008.

Further Reading

Angrist, Joshua D., and Pischke, J-S. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press, 2008.

Morgan, S. L., and Winship, C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* New York: Cambridge University Press, 2007.

Rubin, D. B. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association*, 2005, *100*, 322–331.