5

Quasi-Experimental Designs That Use Both Control Groups and Pretests

Con·trol (kən-trōl′): [Middle English *controllen*, from Anglo-Norman *contreroller*, from Medieval Latin *contrrotulre*, to check by duplicate register, from *contrrotulus*, duplicate register: Latin *contr-*, *contra-* + Latin *rotulus*, roll, diminutive of *rota*, wheel; see *ret-* in Indo-European Roots.] v. tr. con·trolled, con·trol·ling, con·trols. 1. a. To verify or regulate (a scientific experiment) by conducting a parallel experiment or by comparing with another standard. b. To verify (an account, for example) by using a duplicate register for comparison. n. 1. a. A standard of comparison for checking or verifying the results of an experiment. b. An individual or group used as a standard of comparison in a control experiment.

Pre·test (prē′tĕst′): n. 1. a. A preliminary test given to determine whether students are sufficiently prepared for a more advanced course of studies. b. A test taken for practice. 2. The advance testing of something, such as a questionnaire, a product, or an idea. v. tr. and intr. pre·test·ed, pre·test·ing, pre·tests (pr-tst.). To subject to or conduct a pretest.

THE HOMEMAKER–Home Health Aide Demonstration Program provided selected welfare recipients with up to 6 weeks of training, followed by subsidized employment as homemakers and home health aides. To determine if this intervention improved subsequent earnings, Bell et al. (1995) compared results from those who received training with three different nonrandomized control groups: (1) those who applied to the program but left before being screened for eligibility, (2) those who applied but were screened out by staff as ineligible, and (3) those

who applied and were accepted but did not participate in the training.[1] Comparisons of results between treatment and all three control groups suggested that training improved subsequent earnings, although the size of the effect depended on which control group was used. Information about earnings before treatment was available for all these groups, and Bell et al. (1995) showed that those pretest differences were unlikely to account for posttest differences that later emerged.

DESIGNS THAT USE BOTH CONTROL GROUPS AND PRETESTS

This chapter focuses on quasi-experimental designs that, like that of Bell et al. (1995), have both control groups and pretests. The chapter explains how the use of carefully selected comparison groups facilitates causal inference from quasi-experiments, but it also argues that such control groups are of minimal advantage unless they are also accompanied by pretest measures taken on the same outcome variable as the posttest. Such pretests serve many purposes. They tell us about how the groups being compared initially differ and so alert us to the higher probability that some internal validity threats rather than others may be operating. They also tell us something about the magnitude of initial group differences on the variable that is usually most highly correlated with the outcome. The strong assumption is that the smaller the difference on the pretest, the less is the likelihood of strong initial selection biases on that pretest operating, though, unlike with random assignment, there can be no assumption that unmeasured variables at pretest are unrelated to outcome. And finally, having pretest measures helps enormously with the statistical analysis, especially if the reliability of these measures is known. No single variable will usually do as well as the pretest for these purposes. All these reasons explain why we like pretests and control groups in the widely implementable quasi-experimental designs that we cover in this chapter. Table 5.1 summarizes the quasi-experimental designs we consider.

The Untreated Control Group Design With Dependent Pretest and Posttest Samples

Frequently called the nonequivalent comparison group design, this may be the most common of all quasi-experiments. The initial variant we consider uses a treatment group and an untreated comparison group, with both pretest and

1. This study also included a randomized control, but that is not relevant for present purposes.

TABLE 5.1 Quasi-Experimental Designs That Use Comparison Groups and Pretests

Untreated Control Group Design with Dependent Pretest and Posttest Samples

NR O_1 X O_2

NR O_1 O_2

Untreated Control Group Design with Dependent Pretest and Posttest Samples Using a Double Pretest

NR O_1 O_2 X O_3

NR O_1 O_2 O_3

Untreated Control Group Design with Dependent Pretest and Posttest Samples Using Switching Replications

NR O_1 X O_2 O_3

NR O_1 O_2 X O_3

Untreated Control Group Design with Dependent Pretest and Posttest Samples Using Reversed-Treatment Control Group

NR O_1 X- O_2

NR O_1 X- O_2

Cohort Control Group Design

NR O_1

.....

NR X O_2

Cohort Control Group Design with Pretest from Each Cohort

NR O_1 O_2

.....

NR O_3 X O_4

posttest data gathered on the same units.[2] The latter is what makes the *dependent* samples feature. It is diagrammed:

NR O_1 X O_2

NR O_1 O_2

2. A variation is the regression point displacement design. It uses a posttest, a predictor of posttest scores that is taken prior to treatment (the predictor may be a pretest but often is not), and one treatment unit but many control units; each unit contributes a group mean but not data on individuals within groups (Campbell & Russo, 1999; Trochim & Campbell, 1996). The design can sometimes be useful when a single pretest (or other predictor) and posttest are available from so few treatment units that no other design is feasible. This might occur with administrative records in which data are not reported in a disaggregated way and many control units are available and in clinical contexts in which a treatment is given to a single client but records on many control clients are available.

The joint use of a pretest and a comparison group makes it easier to examine certain threats to validity. Because the groups are nonequivalent by definition, selection bias is presumed to be present. The pretest allows exploration of the possible size and direction of that bias.[3] For example, Carter, Winkler, and Biddle (1987) evaluated the effects of the National Institutes of Health (NIH) Research Career Development Award (RCDA), a program designed to improve the research careers of promising scientists. They found that those who received RCDAs did better than those not receiving them, but those who received them had also done better at pretest by a similar amount. So the final difference may have been due more to initial selection bias than to the effects of RCDA. The use of a pretest also allows examination of the nature of attrition, allowing researchers to describe group differences between who does and does not remain in a study. However, the extent to which the pretest can render selection implausible depends on the size of any selection bias and the role of any unmeasured variables that cause selection and are correlated with the outcome. The absence of pretest differences in a quasi-experiment is never proof that selection bias is absent.

When pretest differences do exist, the possibility increases that selection will combine with other threats additively or interactively. For example, selection-maturation may arise if respondents in one group are growing more experienced, tired, or bored than respondents in another group. To illustrate, suppose a new practice is introduced in a setting in which the average pretest level of performance exceeds the average pretest level in the control setting. If the treatment improves outcome, the posttest difference between groups might be even larger than the pretest difference. But this pattern might also occur if treatment group participants were, say, brighter on average and used their higher aptitude to learn at a faster rate than the controls—the rich get richer, so to speak.

A selection-instrumentation threat can occur when nonequivalent groups begin at different points on the pretest. On many scales, the intervals are unequal, and change is easier to detect at some points than at others (e.g., in its middle rather than at its extremes). On normed achievement test scores, for instance, getting a single item correct can have greater implications for percentile rankings at the extremes of a distribution than at the mean. Thus one item translates into different amounts of percentile change depending on the scale position of the respondent. Selection-instrumentation problems are probably more acute (1) the greater the initial nonequivalence between groups, (2) the greater the pretest-posttest change, and (3) the closer any group means are to one end of the scale, so that ceiling or floor effects occur. Sometimes, clues to the presence of such problems are apparent from inspecting pretest and posttest frequency distributions within each group to see if they are skewed or when group means and variances

3. This is typically done by seeing if groups differ significantly at pretest, but it might be better done using equivalency testing methods (Reichardt & Gollob, 1997; Rogers, Howard, & Vessey, 1993). The latter can be more sensitive to detecting pretest differences, although failure to find differences does *not* prove that groups are equal at pretest because groups may still differ on unobserved variables.

are correlated. Sometimes raw data can be rescaled to reduce such problems, whereas at other times a careful choice must be made to use groups that score close to each other at the middle of a scale.

A third example is selection-regression. In the 1969 Head Start quasi-experiment described in the previous chapter (Cicerelli & Associates, 1969), the treatment group of children who attended Head Start was potentially from a different population than the control group children, who did not attend. Recognizing this possibility, the Head Start researchers selected as matched controls only those controls who had the same sex, race, and kindergarten attendance status as the Head Start children. But this led to the problem of differential regression described in the last chapter.

A fourth problem is selection-history (or local history), the possibility that an event (or events) occurred between pretest and posttest that affected one group more than another. For example, a review of federal programs to improve pregnancy outcome (Shadish & Reis, 1984) found that many studies used the pretest-posttest comparison group design, and results suggested that such programs improved pregnancy outcome. But mothers who were eligible for these programs also were eligible for other programs that can improve pregnancy outcome, including food stamps and various health care programs. So it was impossible to know with confidence whether improvements in pregnancy outcome were caused by treatment or by these other programs.

How the Plausibility of Threats Depends Partly on the Observed Pattern of Outcomes

This list of relevant internal validity threats is daunting. However, the plausibility of a threat is always contextually dependent on the joint characteristics of the design, on extrastudy knowledge about the threats, and on the pattern of observed study results. Therefore, *possible* threats to validity are not always *plausible* ones. For example, maturation processes in children that cause *increased* academic achievement are not plausible explanations for *decreased* achievement. To make this point more generally, we now outline five outcome patterns that are observed with the pretest-posttest comparison group design and show how they render threats to validity more or less plausible. We focus mostly on selection-maturation but occasionally comment on other threats as well.

Outcome 1: Both Groups Grow Apart in the Same Direction. A common pattern of selection-maturation occurs when initially nonequivalent groups grow apart at different average rates in the same direction (Figure 5.1). This pattern has been called a fan-spread model of maturation because the groups grow apart over time like ribs in a fan, from the center out to the edges. Standardizing scores makes the fan spread disappear because the fan spread is a function of measured variances growing systematically over time, and standardization involves dividing scores by their variation and so putting scores at each time point on the same

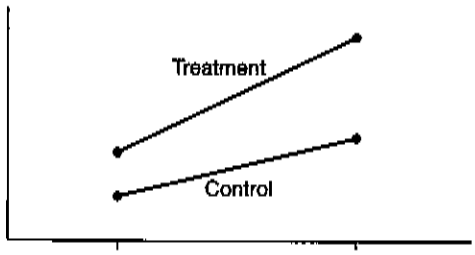

FIGURE 5.1 First outcome of the no-treatment control group design with pretest and posttest

scale instead of on different scales. This pattern is consistent with treatment effects, but can alternative interpretations be identified and ruled out?

Ralston, Anthony, and Gustafson (1985) examined the effects of flexible working hours (flextime) on productivity in two state government agencies. In the agency without flextime, productivity was initially lower and increased slightly over time; in the agency with flextime, it was initially higher but increased at a faster rate. This pattern is common in quasi-experiments, particularly when respondents self-select into condition. But even when administrators assign respondents, treatments are often made available to the especially meritorious, those most keen to improve, or to the more able or better networked, and such persons are also likely to improve at a faster rate for reasons that have nothing to do with treatment.

Several analytic clues can suggest whether nonequivalent groups are maturing at different rates. If group mean differences are a result of this selection-maturation threat, then differential growth *between* groups should also be occurring *within* groups. This could be detected by a within-group analysis in which higher performing members of the group with the higher pretest mean should be growing faster than lower performing members of that same group. This selection-maturation threat is also often associated with posttest within-group variances that are greater than the corresponding pretest variances. It may also help to plot *pretest* scores against the hypothesized maturational variable (e.g., age or years of experience) for the experimental and control groups separately. If the regression lines differ, different growth rates are likely. Such group differences in slope cannot be due to treatment because only the pretest scores have been analyzed.

Nothing makes initial group difference increase linearly; growth can be linear in one condition but quadratic in another. However, in our experience differential maturation of the fan-spread type is commonplace. In education, for example, children who show higher achievement often grow steadily ahead of their lower scoring contemporaries on the original metrics. We suspect that other longitudinal data sets will also show the fan-spread type of differential maturation. Nonetheless, some theoretical formulations predict a different selection-maturation pattern, even in some areas in education. For instance, Piaget's theory predicts sharp discontinuities in growth differences as some children suddenly acquire a concept and

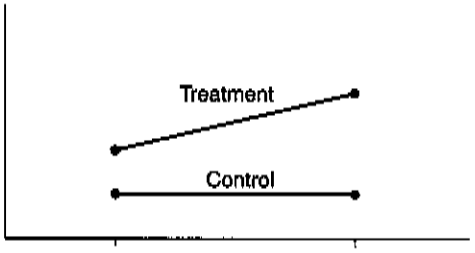

FIGURE 5.2 Second outcome of the no-treatment control group design with pretest and posttest

others do not. So each study using the basic design must present and justify its own assumptions about maturational differences. Sometimes pretest data will play an important role in this. Sometimes data from other longitudinal samples will serve a similar function, as with our assertion that a fan-spread model often fits longitudinal data on academic achievement. But at other times, theoretical speculation is all that can be presented.

Outcome 2: No Change in the Control Group. Narayanan and Nath (1982) used this design to examine how flextime influenced an existing unit of employees compared with another unit in the same company. Results showed improved supervisor-subordinate relations in the flextime group but no changes in the controls, as Figure 5.2 simulates.

When the controls do not change, the critic must explain why spontaneous growth occurred only in the treatment group. It is often easier to think about why both groups mature at different rates in the same direction or why neither group should be changing over time than to think about why one group improves whereas the other does not. Sometimes, within-group analyses can shed light on such between-group threats. For example, if the treatment group was maturing more quickly because the participants were older than those in the control group, then the data could be split by age. If treatment group participants continued to improve no matter what their age, this makes a selection-maturation hypothesis less plausible if it postulates that there should be growth in one group but not the other. Yet when all is said and done, not a lot of reliance can be placed on this particular pattern of differential change. The reason is that it is not unknown for one group to improve and another not to change. Moreover, the pattern of differential change we discussed as more prevalent is only more so in general. Yet each study is highly contextual, and generalities may not apply.

Outcome 3: Initial Pretest Differences Favoring the Treatment Group That Diminish Over Time. Figure 5.3 describes the scenario by which the pretest superiority of a treatment group is diminished or eliminated at posttest. This outcome occurred in a sample of Black third, fourth, and fifth graders in a study

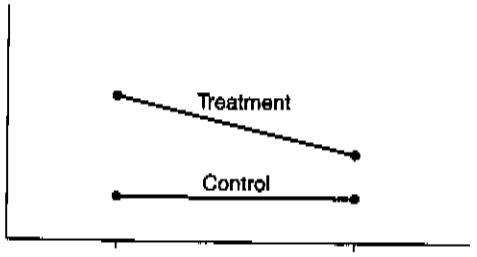

FIGURE 5.3 Third outcome of the no-treatment control group design with pretest and posttest

of the effects of school integration on academic self-concept (Weber, Cook, & Campbell, 1971). At pretest, Black children attending all-Black schools had higher academic self-concept than did Black children attending integrated schools. After formal school integration occurred, the initial difference was no longer found.

Some of the internal validity threats described for Figures 5.1 and 5.2 are also relevant to Figure 5.3. However, selection-maturation is less plausible, for it is rare that those who start off further ahead fall back later on or that those who start further behind subsequently catch up. It can happen, of course. For example, if in an educational context one group were slightly older than another but less intelligent, the older group might be further ahead at the earlier point due to their age advantage but lose this advantage as the younger but smarter group comes to perform better. But such phenomena are rare, and in the Weber et al. (1971) example, the two groups were equivalent in age. Thus the argument is that no presently known maturation process can account for the pattern of results in Figure 5.3, although some such process might be found in the future.

Outcome 4: Initial Pretest Differences Favoring the Control Group That Diminish Over Time. In this case, as in Figure 5.3, the experimental-control difference is greater at pretest than at posttest, but now the experimental group initially *under*performs the controls (Figure 5.4). This is the outcome desired when schools introduce compensatory inputs to increase the performance of the disadvantaged or when a firm makes changes to try to improve a unit's poor performance. Keller and Holland (1981) found this pattern when they assessed the impact of a job change on employee performance, innovativeness, satisfaction, and integration in three research and development organizations. Employees who were promoted or assigned to a different job were the treatment group, and all others were controls. Outcomes were measured twice, 1 year apart. Although this work had no explicit compensatory focus, the data fit the pattern under discussion, and those with the job change showed improved outcomes, whereas the outcomes for others stayed the same.

The outcome is subject to typical scaling (i.e., selection-instrumentation) and local history (i.e., selection-history) threats. But two special elements stand out. First, if the company changed the jobs of those employees whose performance was

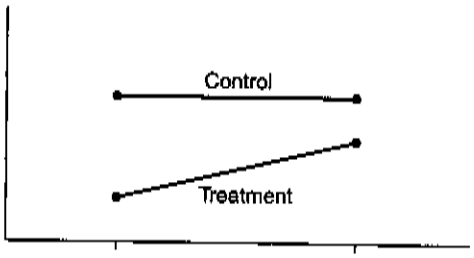

FIGURE 5.4 Fourth outcome of the no-treatment control group design with pretest and posttest

particularly poor at pretest, the outcome for those employees should regress upward at posttest, an outcome that could produce the results in Figure 5.4. If the treatment-control differences in Keller and Holland (1981) were temporally stable, something we could not tell with this design (but could if two pretests were used), then regression would not be a threat. In nonequivalent control group designs, therefore, it is imperative to explore the reasons for initial group differences, including why some groups assign themselves or are assigned to one treatment rather than to another.

The second special element of this design is that the outcome in Figure 5.4 rules out selection-maturation of the fan-spread type—or shows that the treatment overcame such an effect if there were one. However, other selection-maturation patterns could be invoked. In Keller and Holland (1981), for example, the job changers may have been junior staff members in the organization, accounting for their lower pretest scores, but they may also have been particularly open to learning from new experiences, making their performance rise disproportionately quickly. Data on age and time in the organization would have to be analyzed for this possibility. In general, this outcome is often interpretable causally. But it has to be explored seriously in any single study in case its specifics set up complex selection-maturation patterns like those just elaborated.

Outcome 5: Outcomes That Cross Over in the Direction of Relationships. In the hypothetical outcome of Figure 5.5, the trend lines cross over and the means are reliably different in one direction at pretest and in the opposite direction at posttest. This outcome is particularly amenable to causal interpretation. First, the plausibility of selection-instrumentation is reduced, for no simple data transformation can remove the interaction. For example, a ceiling effect cannot explain how the lower-scoring group came to draw ahead of a group that initially scored higher than it did. A convincing scaling artifact would have to postulate that the posttest mean of the treatment group is inflated because the interval properties of the test make change easier at points on the scale that are further away from its mean. However, this explains the exacerbation of a true effect and not the creation of a totally artifactual one.

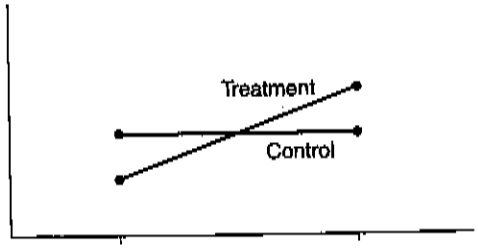

FIGURE 5.5 Fifth outcome of the no-treatment control group design with pretest and posttest

Second, selection-maturation threats are less likely with Figure 5.5, for crossover interaction maturation patterns are not widely expected, although they do occur. An example of the pattern in Figure 5.5 is Cook et al.'s (1975) reanalysis of Educational Testing Service (ETS) data on the effectiveness of *Sesame Street.* They found that children who were encouraged to view the show knew reliably less at pretest than children who were not encouraged to watch but that they knew reliably more than the control group at posttest. But were the encouraged children younger and brighter, thus scoring lower than controls at the pretest but changing more over time because of their greater ability? Fortunately, data indicated that the encouraged and nonencouraged groups did not differ in age or on several pretest measures of ability, reducing the plausibility of this threat.

Third, the outcome in Figure 5.5 renders a regression threat unlikely. Greene and Podsakoff (1978) found the depicted crossover when they examined how removing a pay incentive plan affected employee satisfaction in a paper mill. The employees were divided into high, middle, and low performers, and satisfaction was measured before and after removal of the pay incentive. Following removal, the high performers' satisfaction decreased reliably, that of the low performers increased, and that of the midlevel performers did not change. These slope differences might be due to regression if all three groups converged on the same grand mean (similar to Figure 5.4). But statistical regression cannot explain why the low performers reliably surpassed the high performers at posttest, though regression may have inflated treatment estimates.

Unfortunately, any attempt to set up a design to achieve the outcome shown in Figure 5.5 involves considerable risk. One reason is that the power to detect a statistically reliable interaction is low (Aiken & West, 1991). So such studies must be designed carefully. This is especially true when a fan-spread process such as that shown in Figure 5.2 is expected, for then a no-difference finding would leave it unclear whether the treatment had no effect or whether two countervailing forces (the treatment and fan-spread maturation) had canceled each other. Even if there were a difference in slopes, it would probably take the form of Figure 5.4, not Figure 5.5, and Figure 5.4 is less interpretable. So researchers should not rely on designing research to get the outcome in Figure 5.5. Instead, steps should be taken to add stronger design controls to the basic pretest-posttest design with control group.

Ways to Improve the Untreated Control Group Design With Dependent Pretest and Posttest Samples

As with the designs in the previous chapter, this basic design can be improved substantially by adding thoughtfully chosen design features to address threats to validity that are plausible in the context of the experiment. Examples include the following.

Using a Double Pretest. Here, the same pretest is administered at two different time points, preferably with the same time delay as between the second pretest and the posttest. The design is diagrammed as:

NR	O_1	O_2	X	O_3
NR	O_1	O_2		O_3

The double pretest allows the researcher to understand possible biases in the main treatment analysis—if "treatment effects" emerge in the analysis of O_1 to O_2, similar biases may exist in the analysis from O_2 to O_3. Wortman, Reichardt, and St. Pierre (1978) used this design to study how the Alum Rock educational voucher experiment affected reading test scores. In this program, parents selected a local school for their child and received a voucher equal to the cost of education at that school. The aim was to foster competition between schools in the system. Initial data analysis by others had claimed that vouchers decreased academic performance, but Wortman and colleagues doubted the conclusion that had been drawn. So they followed a group of students through the first to the third grades in both voucher and nonvoucher schools and reanalyzed test scores using a double pretest. Furthermore, they divided the voucher schools into those with and without traditional voucher programs. The additional pretest allowed them to contrast pretreatment growth rates in reading (between O_1 and O_2) with posttest change in rates (between O_2 and O_3), and, because of this, the decrease in reading previously attributed to voucher schools was then attributed only to the nontraditional voucher group. The traditional voucher and nonvoucher groups showed no differential effect that could not be explained by the continuation of the same maturation rates that had previously characterized the traditional and voucher control schools.

The double pretest permits assessment of a selection-maturation threat on the assumption that the rates between O_1 and O_2 will continue between O_2 and O_3. That assumption is testable only for the untreated group. Moreover, the within-group growth rates will be fallibly estimated, given measurement error; and instrumentation shifts could make measured growth between O_1 and O_2 unlike that between O_2 and O_3. So the double pretest design with nonequivalent groups is not perfect. Yet the second pretest can help considerably in assessing the plausibility of selection-maturation by describing the pre-treatment growth differences. The double pretest also helps reveal regression effects if the O_2 observation in either group is atypically low or high compared with O_1. It further helps estimate

more precisely the correlation between observations at different times, something of great value in the statistical analysis. Without the extra time point, the correlation between O_2 and O_3 in the treated group gives an unclear estimate of what the correlation would have been in the absence of a treatment.

Why are multiple pretests not used more often? Ignorance is surely one reason, but another reason is that it is sometimes infeasible. Often one is lucky to be able to delay treatment long enough to obtain a single pretest, let alone two, and let alone being able to space the pretests with the same time interval between pretest and posttest. Sometimes, archives will make possible a second or even more pretests, thus moving toward an even more powerful time series design. In addition, persons responsible for authorizing research expenditures are sometimes loath to see money spent for design elements other than posttest measures. Convincing them about the value of pretests and conventional control groups is hard enough. Convincing them of the value of double pretests can be even harder! Nonetheless, whenever the archival system, time frame, resources, and politics permit, the same pretest should be administered twice prior to treatment.

Using Switching Replications. With switching replications, the researcher administers treatment at a later date to the group that initially served as a no-treatment control. The resulting design is diagrammed as:

NR	O_1	*X*	O_2		O_3
NR	O_1		O_2	*X*	O_3

Besadur, Graen, and Scandura (1986) used a version of this design to study how training affected engineers' attitudes toward divergent thinking in solving problems. Measurement was taken prior to training, following the training of one group of engineers, and then following the training of a second nonequivalent group. The latter group served as controls in the first phase of the study, whereas the roles were switched in the second phase. However, the second phase is not an exact replication. The context surrounding the second treatment is different from the first, both historically and because the treatment has been removed from the first group. Even if the treatment was not removed, it is assumed to have no current impact. (However, the design is still useful even if the initial treatment continues to have an impact, especially if the control group catches up to the treatment group once the control group receives treatment.) Given the contextual differences between the first and second treatment, the second introduction of the treatment is a modified replication, probing both internal validity and an external validity issue of whether this new context changes the treatment effect.

The design can be extended to more groups than two. When it is, it is sometimes possible to assign groups at random to the particular time at which they start treatment, because by definition there must be many consecutively staggered times available if the design is to be implemented with many groups. This random com-

ponent can help strengthen inferences, the more so when many groups at many time points are available. But even without the random assignment of treatments to time intervals, the analytic possibilities are productively expanded when more groups and time points are in the design (e.g., Koehler & Levin, 1998).

The major limitations of this design follow from the fact that later instances of groups serving as controls entail either (1) keeping the same treatment in place but presuming it to have no long-term discontinuous effects in the same direction as the treatment later applied to the initial controls or (2) removing the treatment from the original treatment group. This potentially sets up processes of compensatory rivalry and the like that must be thoroughly described, measured, and used in the analysis. Otherwise, the switching replications design is strong. Only a pattern of historical changes that mimics the time sequence of the treatment introductions can serve as an alternative interpretation.

Using a Reversed-Treatment Control Group. We diagram this version of the design as:

$$\begin{array}{llll} NR & O_1 & X_+ & O_2 \\ \hline NR & O_1 & X_- & O_2 \end{array}$$

where X_+ represents a treatment expected to produce an effect in one direction and X_- represents a conceptually opposite treatment expected to *reverse* the effect. Hackman, Pearce, and Wolfe (1978) used the design to investigate how changes in the motivational properties of jobs affect worker attitudes and behaviors. As a result of a technological innovation, clerical jobs in a bank were changed to make the work on some units more complex and challenging (X_+) but to make work on other units less so (X_-). These changes were made without the company personnel being told of their possible motivational consequences, and measures of job characteristics, employee attitudes, and work behaviors were taken before and after the jobs were redesigned. If treatment X_+ improved the scores of the treatment group, and if treatment X_- decreased the scores of the comparison group, a statistical interaction should result, suggesting a treatment effect.

The reversed-treatment design can have a special construct validity advantage. The causal construct must be rigorously specified and manipulated to create a sensitive test in which one version of the cause (job enrichment) affects one group one way, whereas its conceptual opposite (job impoverishment) affects another group the opposite way. To understand this better, consider what would have happened had Hackman et al. (1978) used an enriched-job group only and no-treatment controls. A steeper pretest-posttest slope in the enriched condition could then be attributed to either the job changes or to respondents feeling specially treated or guessing the hypothesis. The plausibility of such alternatives is lessened in this design if the expected pretest-posttest decrease in job satisfaction is found in the reversed-treatment group because awareness of being in research

is typically thought to elicit socially desirable responses. To explain both an increase in the enriched group and a decrease in the reversed group, each set of respondents would have to guess the hypothesis and want to corroborate it in their own different way.

Interpretation of this design depends on producing two effects with opposite signs. It therefore assumes that little historical or motivational change would otherwise be taking place. When change is differential across treatments but in the same direction, results are less interpretable, because their relationship to a no-treatment control group is unknown. Adding such a control is helpful and should be done when it is feasible. Also in many contexts, ethical and practical considerations prevent using a reversed treatment. Most treatments have ameliorative and prosocial goals, but a conceptually opposite treatment might be harmful. However, that is not clearly the case with Hackman et al. (1978). Who is to say whether it is more beneficial to have one's job made more or less complex than it used to be?

Direct Measurement of Threats to Validity. These measurements allow the researcher to diagnose the possible presence of threats to validity. In Narayanan and Nath (1982), flextime was initiated in one unit of a company while another served as a no-treatment control. However, a history threat could be posed if supervisory practices changed in one group but not the other during the study. To explore this threat, Narayanan and Nath measured such changes and found none. Of course, this is only one example of history, and many others could be discovered, so researchers have to be vigilant lest finding that one study-specific threat is implausible lulls them into believing that all threats are implausible. Each individual threat has to be conceptualized, validly measured, and validly analyzed, making direct measurement of threats difficult. Still, measuring threats can facilitate later statistical analysis by allowing alternative interpretations to be built into whatever analyses are used to deal with initial group nonequivalence.

Matching Through Cohort Controls

Many institutions experience regular turnover as one group "graduates" to another level and their place is taken by another group. Schools are an obvious example, as most children are promoted from one grade to the next each year. Other examples include businesses in which one group of trainees follows another, families in which one sibling follows another, and prisons in which one group of inmates follows another. The term **cohort** designates the successive groups that go through processes such as these.[4] Cohorts are particularly useful as control groups

4. The term *cohort* is used in some other areas (e.g., developmental and longitudinal studies) to refer to any group that is repeatedly measured over time, a very different use from the present one.

if (1) one cohort experiences a given treatment and earlier or later cohorts do not; (2) cohorts differ in only minor ways from their contiguous cohorts; (3) organizations insist that a treatment be given to everybody, thus precluding simultaneous controls and making possible only historical controls; and (4) an organization's archival records can be used for constructing and then comparing cohorts.

The crucial assumption with cohorts is that selection differences are smaller between cohorts than would be the case between noncohort comparison groups. However, this assumption must be probed in each study through, for example, analyses of background characteristics presumed to correlate with outcomes. Even then, presumed comparability will never be as high with cohorts as with random assignment. Further, a review of behavioral genetics research found that, in the area of intellectual performance, environmental differences in the microworlds that siblings live in or create for themselves make two children from the same family as different from one another as are children paired randomly from the population (Plomin & Daniels, 1987). If this conclusion is true and generalizable to nonintellectual domains, it would seriously undermine the case for assigning special status to siblings as cohort controls. Yet many economists include sibling control designs among their preferred armamentarium for studying the effects of external variables on labor force participation or educational attainment (Aronson, 1998; Ashenfelter & Krueger, 1994; Currie & Duncan, 1995, 1999; Duncan, Yeung, Brooks-Gunn, & Smith, 1998; Geronimus & Korenman, 1992).

An example of the use of sibling controls is provided by Minton (1975). She examined how the first season of *Sesame Street* affected Metropolitan Readiness Test (MRT) scores of a heterogeneous sample of kindergarten children. She located a kindergarten in which the test was administered at the end of the child's first year. For a control group, she used MRT scores of the children's older siblings, who had attended the same kindergarten before *Sesame Street* began. So she had the scores from a time at which those siblings were at the same age and maturational stage as their siblings were during the run of *Sesame Street*. The design is diagrammed here; the dotted line (.......) between nonequivalent groups indicates a cohort control. We introduce a cohort design without pretest first and then add a pretest in the next section. The numerical subscripts refer to time of measurement, with the effect assessed by contrasting O_1 to O_2. The design clearly shows that the older sibling group is being used as the same-age, same-maturational-status, reduced-selection control group.

$$\begin{array}{llll} NR & O_1 & & \\ \hdashline NR & & X & O_2 \end{array}$$

Despite similarities between cohorts in maturation status and other family-based variables, to contrast just these two observations provides a weak test of the causal hypothesis. First, a selection problem remains because older siblings are more likely to be first-borns and first-borns tend to outperform later siblings on cognitive achievement tests (Zajonc & Markus, 1975). One way to reduce this

threat is to analyze the data separately by birth order of the older child, because the birth-order effect should dampen as the birth order of the older sibling increases (Zajonc & Markus, 1975). The design is also weak with respect to history, for older and younger siblings could have experienced differential events other than watching *Sesame Street* that affected knowledge levels. One way to explore this threat is to break cohorts down[5] into those whose kindergarten experience was separated by 1, 2, 3, or more years from their siblings to see if the greater learning of the younger group held over the different sets of historical events that these cohorts presumably experienced. But even so, this procedure would still not control for those historical events that took place during the same year that *Sesame Street* was introduced. So a better solution would be to repeat the experiment in different schools in different years. If the effect occurred each time, any historical event or events that masqueraded as treatment effects would have to temporally mimic the introduction of the treatment from school to school across different years. As it turned out, not even this last possibility was feasible with *Sesame Street*, given its great initial popularity in private homes. Hence no school group with minimal exposure would have been possible.

Direct measurement can sometimes help assess selection and history. For instance, Devine, O'Connor, Cook, and Curtin (1990) conducted a quasi-experiment to examine how a psychoeducational care workshop influenced nurses' care of cholecystectomy (gallbladder) surgery patients and their recovery from surgery. Reports were collected from all relevant patients in a single hospital for 7 months before treatment and on another group at the same hospital for 6 months after treatment, thus creating pretreatment and posttreatment cohorts. An analysis of many background characteristics and hospital records revealed no differences between the two cohorts, minimizing the selection threat for the variables examined (but not for unmeasured attributes). Still, it would have been better if circumstances had allowed collecting both pretest and posttest data for a calendar year each instead of for 7 and 6 months, respectively, because the data collection procedure actually implemented is confounded with seasons. Regarding history, the research staff were in the target hospital most days and detected no major irrelevant changes that might have influenced recovery from surgery. This provides no guarantee, of course, and design modifications are better than measurement for ruling out this internal validity threat. So data were also collected from a nearby control hospital that was owned by the same corporation and

5. Such partitioning needs to be done with great caution, especially if it creates more extreme and less extreme groups. Our previous work gave an example of partitioning from the Minton study in which the treatment group was partitioned into four groups by level of viewing of *Sesame Street*. The sibling cohorts were then matched to the same partition. However, Mark (1986, p. 60) identified a plausible regression artifact that may have resulted from this partitioning: "Younger siblings who self-select into heavy 'Sesame Street' viewership are likely to be highly interested in learning, while those who self-select into light viewership are not. Given the less than perfect relationship between sibling's academic skills, we would expect that the older siblings would display less extreme behavior. The result of this regression effect would be a statistical interaction of the sort presented by Cook and Campbell (1979, p. 129) as an 'interpretable outcome.'"

had some of the same physicians. That control also supported the conclusion that the treatment effect was not due to history. What we see with this last point is important, with the cohort design being supplemented by a design feature such as a no-treatment control group. That kind of design improvement is what we now turn to, adding even more design features to improve causal inference.

Improving Cohort Controls by Adding Pretests

In a study comparing the effectiveness of regular teachers and outside contractors hired to stimulate children's achievement, Saretsky (1972) noted that the teachers made special efforts and performed better than would have been expected given their previous years' performances. He attributed this compensatory rivalry to teacher fear of losing their jobs if contractors outperformed them. Assume for pedagogic purposes that he compared the average gain in classes taught by teachers during the study period with the average gain from the same classes taught by the same teachers in previous years. The resulting design would be of the following form, with O_1 and O_2 representing beginning and end of year scores for the earlier cohort, who could not have been influenced by teacher fears, and O_3 and O_4 representing scores for the later cohort that might have been so influenced. The null hypothesis is that the change in one cohort equals that in the other. This design can be extended back over time to include multiple "control" cohorts rather than just one. Indeed, Saretsky reported data for 2 preexperimental years. In principle, if treatment is ongoing, the design could also be extended forward for several years to get multiple estimates of effects.

NR	O_1	O_2			
.....					
NR			O_3	X	O_4

As depicted, the design is similar to the basic nonequivalent control group design with pretest and posttest. The major differences are that measurement occurs at an earlier time period in the control group and that cohorts are assumed to be less nonequivalent than most other nonmatched groups would be. This last point can be explored by comparing cohort pretest means, one of the major advantages of including pretests in cohort design. The pretest also increases statistical power by allowing use of within-subject error terms. It enables better assessment of maturation and regression, and it enters into better (but still imperfect) statistical adjustment for group nonequivalence.

History is a salient internal validity threat in this design—it can involve any event correlated with the outcome that appears only during the O_3–O_4 period, even if there is a series of cohort control periods. Only if a nonequivalent control group is added to the design and measured at exactly the same time points as the treatment cohorts can we hope to address history. Sometimes the design can be strengthened by adding nonequivalent dependent variables if these are appropriate for the topic under investigation.

A variant of this design is what Campbell and Stanley (1963) called the *recurrent institutional cycle* design. With access to school records, or having at least 2 years to do a study with original data collection, the design is:

NR	X	O_1			
NR			O_2	X	O_3
NR					O_4

It involves the three cohorts entering, say, the second grade in 3 consecutive years. The first receives the treatment and a posttest, the second the treatment with both pretest and posttest, and the third no treatment and only one assessment. Note that O_1 and O_2 might not be simultaneously observed because one might be at the end of a school year and the other at the beginning of the next. This cycle is repeated again with O_3 and O_4. A treatment main effect is suggested by a certain pattern of results—that is, if O_1 and O_3 are, say, higher than O_2 and O_4; if O_2 does not differ from O_4; and if O_1 does not differ from O_3. A partial control for history is provided if, in addition to O_3 being greater than O_2, O_1 surpasses O_2, and O_3 surpasses O_4. Then there is presumptive evidence that the treatment could have been effective at two different times, though it is also possible for two separate historical forces to have operated or for one historical force to have reoccurred. But still, any single history alternative would have to repeat to explain both $O_1 > O_2$ and $O_3 > O_4$. Selection is also reduced in this version of a cohort design when the same persons are involved in the O_2–O_3 comparison.

Another threat, testing, is possible because some comparisons involve contrasting a first testing with a second testing (O_2 to O_3). Hence, Campbell and Stanley (1963) recommended splitting the group that is both pretested and posttested into random halves, one of which receives a pretest but the other of which does not. A reliable difference between these two groups at posttest might be due to testing; the lack of such differences would suggest that testing effects are not a problem. Finally, because causal interpretation depends on a complex pattern of outcomes in which three contrasts involve O_2, a change in elevation of O_2 would have crucial implications. Hence the design should be used only with reliable measures and large samples.

Improving Cohort Designs With a Nonequivalent Dependent Variable

Minton (1975) used a nonequivalent dependent variable to improve her study of how the first season of *Sesame Street* affected kindergarten children's learning. She showed, for those who watched *Sesame Street,* that their knowledge of letters that were taught on *Sesame Street* improved significantly more than did their knowledge of letters that were not taught. This outcome helped address maturation threats to validity, because children typically grow in their knowledge of letters of the alpha-

bet over time as a result of many influences, including their own cognitive development. If only maturation explained the results, then we would expect no difference between knowledge of letters that were taught versus those that were not.

DESIGNS THAT COMBINE MANY DESIGN ELEMENTS

Throughout this chapter we have emphasized the value of adding design elements to aid causal inference. In this section, we describe three examples of designs that use many elements, examples that serve to clarify and extend the underlying rationale.

Untreated Matched Controls With Multiple Pretests and Posttests, Nonequivalent Dependent Variables, and Removed and Repeated Treatments

In an exemplar of good quasi-experimental design, Reynolds and West (1987) assessed the effects of Arizona's "Ask for the Sale" campaign to sell lottery tickets. Participating stores selling lottery tickets agreed to post a sign reading, "Did we ask you if you want a Lottery ticket? If not, you get one free," and they also agreed to give a free ticket to those customers who were not asked if they wanted one but who then requested one. Because participation was voluntary, the resulting nonequivalent control group design was supplemented in four ways. First, the authors matched treatment stores to control stores from the same chain (and when possible, from the same zip code area), as well as on the pretest market share of ticket sales. Second, they added multiple pretest and posttest assessments by examining mean weekly ticket sales for 4 weeks before and 4 weeks after the treatment started. Pretest sales trends were decreasing nearly identically in both the treatment and control groups, so that maturation differences could not explain increasing ticket sales. Similarly, regression to the mean was unlikely because the treatment group sales were *continuously* decreasing over four consecutive pretests and because control group ticket sales continued to decrease after treatment began. Third, Aiken and West studied treatment effects on three nonequivalent dependent variables in the treatment group, discovering that the intervention increased ticket sales but not sales of gas, cigarettes, or grocery items. Fourth, they located some stores in which the treatment was removed and then repeated or was initiated later than in other stores and found that the outcome tracked the introduction, removal, and reinstatement of treatment over time whereas sales in the matched controls remained unchanged. Nearly all these analyses suggested that the "Ask for the Sale" intervention increased ticket sales after the program began, making it difficult to think of an alternative explanation for the effect.

Combining Switching Replications With a Nonequivalent Control Group Design

Sometimes the researcher can introduce treatment to part of the original control group, with other controls remaining untreated over this later time period. Sometimes the researcher can even reintroduce treatment a second time to some of the original treatment group to evaluate the benefits of additional treatment. Gunn, Iverson, and Katz (1985) did this in a study of a health education program introduced into 1,071 classrooms nationwide. The design is diagrammed as follows, with *R* indicating a potential use of random assignment that is a useful but not necessary adjunct to the design:

	Year 1			Year 2			
NR	O_1	*X*	O_2	*R*	O_3	*X*	O_4
				R	O_3		O_4
NR	O_1		O_2	*R*	O_3	*X*	O_4
				R	O_3		O_4

Classrooms were first divided into nonequivalent treatment and control groups. Students in each group were tested on knowledge of health before and after the first year of the program. Then the initial control group was divided randomly in half. One half received the health education program to replicate the treatment effect, and the other remained without instruction. In addition, a random sample of the original treatment group received a second year of instruction to explore the incremental benefit of additional health education. Here we see switching replications yoked to a continuation of the original controls and to a treatment booster. This yoking strengthens a switching replications design, especially if the second phase of the study uses random assignment or if those receiving the booster session are identified by falling to one side of a cutoff on a measure of need for the booster session—a regression discontinuity design (see Chapter 7).

An Untreated Control Group With a Double Pretest and Both Independent and Dependent Samples

To evaluate community-level interventions designed to reduce cardiovascular risk factors, both Blackburn et al. (1984) and Farquhar et al. (1990) combined a double pretest with samples in which the outcome was measured on both independent and dependent samples. We diagram the logic of the design here, using per-

pendicular lines between O's to show independent samples and glossing over some complexities in the actual designs used in these two studies.

R	O_1	|	O_2	|	O_3	|	O_4	|	O_5
R	O_1	|	O_2	| X	O_3	|	O_4	|	O_5
R	O_1		O_2		O_3		O_4		O_5
R	O_1		O_2	X	O_3		O_4		O_5

The first two rows of this diagram portray a randomized experiment with communities being assigned to intervention or control and with a cross-sectional panel survey being administered to independent samples of community households at each annual time point (however, relatively few communities were used in each study, and so there can be no presumption here that much initial equivalence was achieved). The two bottom rows of the diagram depict a longitudinal survey of respondents who were followed over time. The major study outcomes were annual physiological measures of heart problems, including blood pressure and cholesterol level. In the cross-sectional panel survey, independent random samples were drawn, both out of concern that obtrusive annual physiological measurement would sensitize repeatedly measured respondents to the treatment and out of desire to generalize to the community at large as it spontaneously changed over time. Because there were only three matched communities in Blackburn's study and two in Farquhar's, the double pretest was used to estimate preintervention linear trends. However, in the Blackburn study, variability between years within cities was greater than expected, and statistical adjustments for this were not very helpful. So Blackburn modified the design in midstream so that some pretest respondents were followed up at several posttests, thus creating the longitudinal sample to complement the independent samples that continued to be drawn. The Farquhar study was designed from scratch to include both independent and dependent samples.

The use of so many different design elements provided many ways to examine the threats to validity (Chaffee, Roser, & Flora, 1989). For example, Chaffee et al. examined history by comparing differences between successive waves of independent samples in the control cities, and they examined attrition by comparing differences between the dependent and independent treatment group samples with differences between the corresponding control group samples. The combined effects of testing and maturation are suggested by comparing the differences between changes over time in the dependent samples (in which testing and maturation are more likely to occur) with changes in the independent samples (although some maturation of the entire population might also occur in these). None of these ways of examining threats is perfect, each providing suggestive rather than definitive evidence.

The Farquhar study is interesting for another reason that is relevant when the unit of assignment is a large aggregate such as a community or a business. For reasons of cost and logistics, it is rarely possible to have many such aggregates. Indeed, the publication of the Farquhar study reported only two treatment and two control communities. Cardiovascular disease decreased in the two treatment communities and in one of the control communities by amounts that hardly differed. But the risk appears to have increased over time in the second control community, despite a national trend downward over the period studied. Omitting this one community from some analyses would have reduced the treatment-control differences to nearly zero. With so few units, there is no pretense of achieving comparability between treatment and control groups, however conscientiously communities were paired before assignment. To deal with this problem requires adding more communities (which will often be prohibitively expensive) or combining studies with similar treatments. In this last case, the treatments will not be identical, and other contextual and evaluation factors will surely also differ between the studies. There is no compelling reason why there should be as many control as experimental units, so adding more control communities is sometimes inexpensive and can increase power (Kish, 1987).

THE ELEMENTS OF DESIGN

We have shown how even the weakest quasi-experimental designs can be strengthened by adding thoughtfully chosen design elements that reduce the number and plausibility of internal validity threats. Here we summarize those design elements (Table 5.2). After all, quasi-experiments are nothing more than combinations of such elements selected to suit particular circumstances of research (Corrin & Cook, 1998). For convenience, we place them into four groups having to do with (1) assignment, (2) measurement, (3) comparison groups, and (4) treatments.

Assignment

In most quasi-experiments, assignment is not controlled by the researcher. Rather, participants self-select into conditions, or someone else makes the assignment decision, as when a physician decides who will receive surgery or a teacher or school board decides which student or school should receive new resources. There is considerable evidence that *nonrandom assignment* often (but not always) yields different results than random assignment does (Chalmers et al., 1983; Colditz, Miller, & Mosteller, 1988; Lipsey & Wilson, 1993; Mosteller, Gilbert, & McPeek, 1980; Wortman, 1992), more so when participants self-select into conditions than when others make the selection decision (Heinsman & Shadish, 1996; Shadish, Matt, Navarro, & Phillips, 2000; Shadish & Ragsdale, 1996)—so self-selection should be avoided if

TABLE 5.2 Design Elements Used in Constructing Experiments and Quasi-Experiments

Assignment
- Random Assignment
- Cutoff-Based Assignment
- Other Nonrandom Assignment
- Matching and Stratifying
- Masking

Measurement
- Posttest Observations
  - Single Posttests
  - Nonequivalent Dependent Variables
  - Multiple Substantive Posttests
- Pretest Observations
  - Single Pretest
  - Retrospective Pretest
  - Proxy Pretest
  - Repeated Pretests Over Time
  - Pretests on Independent Samples
- Moderator Variable with Predicted Interaction
- Measuring Threats to Validity

Comparison Groups
- Single Nonequivalent Groups
- Multiple Nonequivalent Groups
- Cohorts
- Internal Versus External Controls
- Constructed Contrasts
  - Regression Extrapolation Contrasts
  - Normed Contrasts
  - Secondary Data Contrasts

Treatment
- Switching Replications
- Reversed Treatments
- Removed Treatments
- Repeated Treatments

possible. Certain nonrandom assignment methods such as alternating assignment can sometimes approximate random assignment decently well (McAweeney & Klockars, 1998; Staines, McKendrick, Perlis, Sacks, & DeLeon, 1999).

Assignment can often be controlled in other ways than by random methods. *Matching* and *stratifying* can both increase group similarity. However, matching requires significantly more vigilance in quasi-experiments than in randomized

experiments, for when done with unreliable, single measures at one point in time it can create more problems than it solves. When feasible, *masking* (blinding) of investigators, participants, or other research and service staff to assignment can be useful. It prevents two biases: (1) investigator and participant reactivity to knowledge of the condition to which the participant has been assigned and (2) efforts by those involved in assignment to influence results from the condition to which a participant is assigned. In general, then, not all nonrandom assignments are alike, and nonrandom assignments can be improved by preventing self-selection and by using other experimental controls such as matching and masking in cases in which they are feasible.

Measurement

Researchers can improve causal inference by controlling the nature and scheduling of measurements in a study. The major reason for assessing *posttests* after a treatment is to eliminate ambiguity about the temporal precedence of cause and effect. This threat is most likely to occur when a measure of outcome is taken simultaneously with treatment, as occurs in many correlational studies in which the same questionnaire is used to assess both treatment exposure levels and outcome. It is obviously better to separate temporally the measurement of these two crucial attributes of causal analysis. The special posttest called a *nonequivalent dependent variable* requires posttest measurement of two plausibly related constructs (e.g., two measures of health), one of which (the target outcome variable) is expected to change because of the treatment, whereas the other (the nonequivalent dependent variable) is not predicted to change because of the treatment, though it is expected to respond to some or all of the contextually important internal validity threats in the same way as the target outcome (e.g., both would respond in the same degree to a maturational process that improves health across all health measures). If the target outcome variable changes in response to treatment but the nonequivalent dependent variable does not, the inference that the change is due to the treatment is strengthened. If both change, the inference is weakened because the change could have been due to the threats. The use of *multiple substantive posttests* allows the researcher to examine a pattern of evidence about effects. When this pattern is predicted based on prior knowledge about the pattern typically left by a particular cause, more confident causal inference is possible.

Adding a *pretest* to a design helps examine selection biases and attrition as sources of observed effects. Adding *repeated pretests* of the same construct on consecutive occasions prior to treatment helps reveal maturational trends, detect regression artifacts, and study testing and instrumentation effects. Sometimes when it is not possible to collect pretest information on the outcome variable, *retrospective* pretests ask respondents to recall their pretest status; or *proxy* pretests can be gathered on a variable that is correlated with the outcome. These options

can help clarify selection and attrition biases, though more weakly than can pretests on the outcome variable itself. Or one can sometimes gather pretest information on an *independent pretest sample*—participants different from those in the posttest sample but presumed to be similar to them, such as a random sample from the same population.

A *moderator variable* influences the size or direction of an observed effect. It can aid causal inference when the researcher successfully predicts an interaction between the moderator and treatment in producing the observed effect. This confirmation usually allows few plausible threats to internal validity. Finally, *measuring threats to validity* that can be anticipated at the start of the study helps the researcher to detect the occurrence of the threat and whether its direction mimics the observed outcomes. Measuring the presumed selection process is one particularly crucial example.

Comparison Groups

Comparison groups provide data about the counterfactual inference, that is, about what would have happened in the absence of treatment. In quasi-experiments, the counterfactual inference often depends on a *nonequivalent comparison group* deliberately chosen to have maximum pretest similarity to the treatment group on as many observed characteristics as possible or on some particular feature that the researcher believes will be a particularly salient threat to validity. Using thoughtfully chosen *multiple nonequivalent comparison groups* rather than just one comparison can expand the researcher's ability to explore more threats to the causal inference and to triangulate toward a narrower bracket within which the effect is inferred to lie. A particularly useful comparison is to *cohort controls,* to groups that move through an institution (e.g., a school) in cycles (e.g., a new third-grade class each year). Cohorts are thought to be more comparable to each other (e.g., of the same age, same general socioeconomic status, etc.) than are most other nonequivalent comparison groups.

Nonrandom comparisons to an *internal* rather than an *external control group* can sometimes yield more accurate results (Aiken et al., 1998; Bell et al., 1995; Heinsman & Shadish, 1996; Shadish & Ragsdale, 1996). Internal controls are drawn from the same pool of participants (e.g., from students in the same school or class or from all program applicants). External controls are drawn from patently different pools (e.g., patients in different treatment settings) and are presumed to have less in common. Drawing the line between internal and external controls is sometimes difficult, however, and it is clear that all these nonequivalent comparison groups can yield significant biases (Stewart et al., 1993).

Sometimes counterfactual inferences are supported from less desirable sources, including (1) a *regression extrapolation* in which actual and projected posttest scores are compared, (2) a *normed* comparison in which treatment group

scores are compared with normed samples from test manuals and the like, and (3) a *secondary data* comparison in which treatment respondents are compared with samples drawn from other studies. The usefulness of such comparisons depends on the extent to which similarity to the treatment group can be shown, on whether useful matching is possible, and on whether multiple comparisons can be constructed. In the case of the contrasts listed in this paragraph, it would be rare to discover that they adequately describe the missing counterfactual inference.

Treatment

The researcher's ability to control the application and scheduling of treatment is a powerful tool for facilitating a causal inference. The *switching replication* method replicates the treatment effect at a later date in a group that originally served as a control. Better still is the use of multiple comparison groups that each receive treatment at a different time. The *reversed treatment* method applies a treatment expected to reverse the outcome when compared with the expected outcome in the treatment condition. The *removed treatment* method first presents and then removes treatment to demonstrate that the pattern of outcomes follows the pattern of treatment application; and the *repeated treatments* method reintroduces the treatment after its removal, doing so as often as feasible (sometimes called the ABAB design, with A signifying treatment and B signifying treatment removal).

Design Elements and Ideal Quasi-Experimentation

Is there an ideal or best quasi-experimental design, one that assembles these elements optimally? The answer is "Usually not," because the best design for a given study depends on the particular hypotheses being probed, on the contextual relevance of various threats to inference, on knowledge from prior studies about the viability of those threats, and on what design elements are feasible to include. However, most quasi-experiments have used very few of the potentially available quasi-experimental design elements; and our impression is that most quasi-experiments would have benefited by more attention to both the threats to inference and the design elements that might help reduce the plausibility of those threats.

Our advice is in the spirit of R. A. Fisher, who advised researchers to "Make your theories elaborate" (cited in Rosenbaum, 1984, p. 41) in order to improve causal inference from nonrandomized experiments. It is also in the spirit of Holland (1989), who noted two competing principles in drawing causal inferences from quasi-experiments: (1) causal inference in nonrandomized studies requires more *data* than in randomized studies and (2) causal inference in nonrandomized studies requires more *assumptions* in data analyses than in randomized studies.

Holland encouraged researchers to put more emphasis on the former principle (gathering more data) than the latter (making more assumptions), for gathering more data is often the only way to test the assumptions necessary to make better analyses. Adding more design elements is a way to gather more elaborate and diverse data in the service of improving causal inference.

CONCLUSION

Our review in the previous two chapters has noted that the most frequently used quasi-experimental designs typically support causal conclusions that are somewhat ambiguous. In light of this, users must be prepared to tolerate the ambiguity, assume that alternative causal explanations are negligible, or use stronger designs. This chapter emphasizes building stronger designs through adding design features that reduce the plausibility of validity threats in the context under study. In the next chapter, we continue the same theme. By themselves, interrupted time series provide a particularly strong structure for supporting causal inferences. But when the design features just summarized (e.g., comparison groups, nonequivalent dependent variables, switching replications) are added to the interrupted time series, the result is a quasi-experiment whose inferential yield sometimes rivals that of the randomized experiment.

APPENDIX 5.1: IMPORTANT DEVELOPMENTS IN ANALYZING DATA FROM DESIGNS WITH NONEQUIVALENT GROUPS

Statisticians and economists have recently devoted substantial attention to the analysis of data from designs with nonequivalent groups. Much of it is highly statistical and beyond the scope of our focus on design. However, a chapter on quasi-experimentation would be remiss if it did not introduce these developments that we hope will serve not as alternatives to quality quasi-experimental designs but as adjuncts in dealing with whatever biases the best possible design cannot deal with. In a sense, the motto is "statistical adjustment only after the best possible design controls have been used." Winship and Morgan (1999) provide a superb review of this material.

Propensity Scores and Hidden Bias

Throughout the 20th century, statisticians have preferred randomized experiments, paying less attention to quasi-experiments (Shadish & Cook, 1999). This

preference is partly due to the inherent intractability of selection bias, for it is difficult to develop statistical models when the underlying processes are by their very nature unknown. Recently, however, some statisticians have studied these problems with useful results (e.g., Holland, 1986; Rosenbaum, 1984, 1995a; Rubin, 1974, 1991). Much of that work is summarized by Rosenbaum (1995a); and useful examples now exist in epidemiology (C. Drake & Fisher, 1995), medicine (Connors et al., 1996; Smith, 1997; Stone et al., 1995), the evaluation of job training programs (Dehejia & Wahba, 1999), and high school education (Rosenbaum, 1986), to name a few.

A useful development is the propensity score: the predicted probability of being in the treatment (versus control) group from a logistic regression equation.[6] Careful measurement of likely predictors of selection into groups will improve the accuracy of propensity scores. The goal is to include all variables that play a role in the selection process (including interactions and other nonlinear terms; Rosenbaum & Rubin, 1984; Rubin & Thomas, 1996) and that are presumptively related to outcome, even if only weakly so (Rubin, 1997): "Unless a variable can be excluded because there is a consensus that it is unrelated to outcome or is not a proper covariate, it is advisable to include it in the propensity score model even if it is not statistically significant" (Rubin & Thomas, 1996, p. 253). Sample size allowing, some authors suggest also using as predictors *any* pretest variables that differentiate between nonequivalent groups (Canner, 1984, 1991; Cochran, 1965; Rosenbaum, 1995a; Rubin & Thomas, 1996) at a higher than usual Type I error rate (e.g., $p < .10$ or $p < .25$). Predictors should not be caused by the treatment, which usually entails using measures collected before treatment begins. Data tentatively suggest that correctly modeling the form of the regression (i.e., correct inclusion of interaction or nonlinear terms) is less important than including all the relevant predictors of group membership (Dehejia & Wahba, 1999; C. Drake, 1993). In cases of multiple treatments, propensity scores may be computed separately for each pairwise comparison (Rubin, 1997).

The logistic regression reduces each participant's set of covariates to a single propensity score, thus making it feasible to match or stratify on what are essentially multiple variables simultaneously. Standard matching can be used in which one treatment and one control unit are paired. But Rosenbaum (1995a) shows that such pair-matching usually will not minimize the distance between groups within strata on the propensity score. Instead, he recommends *optimal matching*, in which each subset consists of (1) a single treated participant and one or more controls or (2) a single control participant and one or more treated participants. Optimal matching uses an algorithm for minimizing aggregate sample differences between treatment and control conditions on the propensity score. It allows for eliminating prior matches to create new ones if that procedure yields the lowest total

6. Stone et al. (1995) illustrate an alternative method for creating propensity scores using classification tree algorithms rather than logistic regression.

difference over conditions (Rosenbaum, 1995a). Bergstralh, Kosanke, and Jocobsen (1996) provide a SAS macro for optimal matching, and Isserman and Rephann (1995) present a social science example of its application. Many other variations on matching algorithms are possible (e.g., Dehejia & Wahba, 1999; Gu & Rosenbaum, 1993; Heckman, Ichimura, & Todd, 1997; Marsh, 1998). For example, one can match on propensity scores while simultaneously matching on other variables, such as gender or age (Rosenbaum, in press). There is as yet no thorough review of the advantages and disadvantages of all these matching options.

If stratifying, "Cochran (1968) shows that five subclasses are often sufficient to remove over 90% of the bias due to the subclassifying variable or covariate" (Rosenbaum & Rubin, 1984, p. 516). So five strata are typically constructed that contain all the experimental and control cases that fall within the same quintile on the propensity score. That stratification is not affected by violations of linearity, and it balances treatment and control groups in the sense that within any stratum that is homogeneous in the propensity score, differences between treated and control participants on the predictors will be due to chance if the propensity score stratification worked well. The treatment group mean is then estimated as an unweighted average of the five treatment group strata means, and the control group mean is estimated similarly. Alternatively, Robins, Greenland, and Hu (1999) report a method for weighting proportional to the propensity of receiving the treatment actually received that may have advantages, particularly for time-varying treatments. The researcher should test how well stratification on propensity scores succeeded in adjusting for differences in observed covariates by submitting each covariate (separately) and the propensity score itself to a 2 (treatments) $\times$ 5 (strata) analysis of variance. A significant interaction suggests that the propensity score did not adjust well for observed covariates, a situation that is more likely to occur the more seriously discrepant the two groups are on pretest covariates. Sometimes this problem can be ameliorated by adding nonlinear terms to the propensity score equation.

Finally, the propensity score can be used as a covariate in ANCOVA. When the usual ANCOVA assumptions are met and the model is precisely correct (e.g., it models curvilinearity correctly), covariance adjustment is more efficient than matching or stratifying. However, if the model is not substantially correct, covariance adjustments may fail to reduce overt bias or may even increase it (Rosenbaum, in press). Some authors doubt how well covariance models can model the correct functional form (e.g., H. White, 1981; Winship & Morgan, 1999). Dehejia and Wahba (1999) found that matching performed better than covariance compared with a randomized experiment benchmark despite the addition of some nonlinear terms to the covariance model. Fortunately, matching or stratifying on propensity scores can be used in combination with a subsequent covariance analysis, the result being more efficient and robust than when either is used alone (Rosenbaum, 1998, in press). This ANCOVA may include predictors of group membership that were used to compute the propensity score (Rubin & Thomas, 2000; Stone et al., 1995). Although the latter may seem unusual, a predictor may account for both

variability in group membership and variability in outcome. To the extent that those sources of variability are orthogonal (mostly an empirical question in any given case), including the predictor in the final outcome equation can increase the efficiency and decrease the bias of the final estimates.

Four caveats temper our excitement about the potential of propensity scores. First, they work best with larger samples (Rubin, 1997), but many quasi-experiments have small samples. Second, researchers should inspect the overlap between conditions on propensity scores. When overlap is extremely limited, it does not allow identification of many strata or matches with members from the treatments under contrast, which can severely limit sample size, generalizability, and accuracy of any causal conclusions. Third, methods for computing propensity scores when predictors are missing are just now being explored (e.g., D'Agostino & Rubin, 2000); this issue is crucial in practice, as missing data are common. Fourth, the method assumes that no further unknown confounding variable exists that predicts the propensity to be in condition and that is correlated with outcome. This is a strong assumption. Random assignment balances treatments on both observed and unobserved covariates on expectation; but propensity score adjustments balance treatments only on observed covariates, leaving hidden bias due to unobserved covariates. It helps reduce hidden bias if propensity scores are constructed from as many predictors of group membership and outcome as is contextually feasible. However, it is rarely possible to know all such variables; and cost or logistical constraints often prevent researchers from measuring those that are suspected to be operating. So hidden bias may remain in quasi-experimental estimates of treatment effects even when the best propensity score analysis is used.

A second relevant development in statistics derives directly from the likelihood of this hidden bias. It is the development of sensitivity analyses to assess whether hidden biases of various sizes would change the results of the study. Such analyses explore how much hidden bias would need to be present to change study outcomes, commonly from a significant observed difference between groups to a finding of no difference or vice versa. Rosenbaum (1991a, 1991b) provides a simple example of the computations (see also Gastwirth, 1992; Gastwirth, Krieger, & Rosenbaum, 1994; S. Greenhouse, 1982; Marcus, 1997b; Psaty et al., 1999; Rosenbaum, 1986, 1987, 1988, 1989, 1991a, 1991b, 1993, 1995a, 1995b, 1999b; Rosenbaum & Krieger, 1990). Similar research has emerged recently in the econometrics literature covered in the next section (Manski, 1990; Manski & Nagin, 1998).

The sensitivity analysis outlined by Rosenbaum (1991a, 1991b, 1995a) works as follows. In a randomized experiment using simple random assignment, the odds of being assigned to treatment or control conditions are even, so the probability of being assigned to treatment is .50. In this case, the significance level (i.e., the Type I error rate) yielded by the statistical test for the difference between the two groups is accurate. In nonrandomized experiments, however, these probabilities may depart from .50; for example, males may be more likely than females to be admitted to a job training intervention. As these probabilities depart from 50/50, the significance

level yielded by a statistical test of the difference between groups can become less accurate, assuming that the omitted variable causing the bias is related to outcome. Unfortunately, without knowing the hidden biases that cause this change in assignment probabilities, we cannot know if the significance levels are too low or too high. A sensitivity analysis identifies how far the lowest and highest possible significance levels will depart from what would have been yielded by a randomized experiment. It does this separately for different assumptions about the degree to which the probabilities of being assigned to conditions depart from .50. This analysis can provide important diagnostic information about the degree of assignment bias on a variable related to outcome that would change the significance of the result.

Rosenbaum (1991a) provides an example in which the observed significance level in a quasi-experiment is $p = .0057$, suggesting that treatment is effective. Sensitivity analysis on the raw study data showed that possible significance ranges from a minimum of .0004 to a maximum of .0367 when the probability of being assigned to conditions ranges from .4 to .6. Both this minimum and maximum would support the conclusion that treatment is effective. However, that narrow assignment probability range (40/60) reflects relatively little departure from the randomized experiment due to hidden bias. If the probability of being assigned to conditions ranges from .25 to .75, then the minimum significance level is $<.0001$ but the maximum is .2420, the latter suggesting no significant effect. The probabilities suggest that if unmeasured variables exist that affect assignment to conditions in this study so that some people are more likely than others to be assigned to treatment by a factor of 3:1 (i.e., .75 to .25), then hidden bias may be creating a false treatment effect where none actually exists (or it may be masking even larger treatment effects).

The pattern of minimum and maximum significance levels and the disparities in assignment probabilities required to produce them will vary over studies. Some studies will seem invulnerable to all but the most extreme assumptions about hidden bias, and others will seem vulnerable to nearly any assumptions. However, sensitivity analyses do not actually indicate whether bias is present, only whether a study is vulnerable to biases of different degrees. Rosenbaum (1991a) describes one study that seemed invulnerable to hidden biases that caused assignment probabilities ranging from .09 to .91; but later research showed that an even larger bias probably existed in that study. The actual detection of hidden bias in a study is not easily accomplished; but sometimes the design elements we have outlined in this chapter and the previous one, such as use of nonequivalent dependent variables or of control groups that have known performance on some unobserved covariate, are useful. For example, Dehejia and Wahba (1999) suggest that when propensity score adjustments on multiple nonequivalent comparison groups yield highly variable results, the possible presence of hidden bias is suggested.

When sensitivity analysis is combined with matching on propensity scores, these new statistical developments provide an important new analytic tool in the arsenal of quasi-experimentation. We hope they are used more widely to help us gain more practical experience about their feasibility and accuracy.

Selection Bias Modeling

Given that propensity score analysis cannot adjust for hidden bias and that sensitivity analyses cannot indicate whether such biases are present, it would be desirable to have a method that remedies these weaknesses. For the past 25 years, a number of economists, notably Heckman (e.g., Barnow, Cain, & Goldberger, 1980; Cronbach, Rogosa, Floden, & Price, 1977; Director, 1979; W. Greene, 1985, 1999; Heckman, 1979; Heckman & Hotz, 1989a, 1989b; Heckman, Hotz, & Dabos, 1987; Heckman & Robb, 1985, 1986a[7]; Stromsdorfer & Farkas, 1980), have developed procedures they hoped would adjust for selection biases between nonequivalent groups to obtain an unbiased estimate of treatment effects. These methods are statistically complex and are not always easily implemented by those without advanced statistical training. They comprise a family of models that make different assumptions about selection. Accessible overviews are provided by Achen (1986), Foster and McLanahan (1996), Moffitt (1991), Newhouse and McClellan (1998), Rindskopf (1986), Winship and Mare (1992), and especially Winship and Morgan (1999).

A simple selection bias model might use two equations, a selection equation and an outcome equation. As with propensity score models, the selection equation predicts actual group membership from a set of presumed determinants of selection into conditions, yielding a predicted group membership score. This score may be substituted for the treatment dummy variable in the outcome equation or added to that equation in addition to the dummy variable. If the selection equation predicts group membership nearly perfectly, and if other assumptions, such as normality of observations, are met, then in principle the coefficient associated with the predicted dummy treatment variable in the effect estimation equation can yield an unbiased estimate of the treatment effect. Unlike propensity score methods, selection bias models can allow for correlation of errors in the selection equation and the outcome equation. This correlation is gained at a cost of assuming the nature of the bivariate relationship between the errors, usually as bivariate normal.

These models are closely related to regression discontinuity designs that achieve an unbiased estimate of treatment effects through full knowledge of the selection model by which participants were assigned to conditions, entering that model (i.e., the cutoff variable) directly into the effects estimation model. Regression discontinuity does not require a selection equation because the design forces perfect prediction of selection based on the cutoff score, so the residual of prediction into conditions is zero. In selection bias models, by analogy, if the residual of the selection equation departs much from zero (which is to say that predicted group membership does not match actual group membership well), then the selection bias model may fail to yield unbiased estimates of treatment effects. This primarily hap-

7. Wainer (1986, pp. 57–62, 108–113) reprints a spirited discussion by John Tukey, John Hartigan, and James Heckman of both the 1985 and 1986 versions of the Heckman and Robb papers.

pens if a variable that would improve prediction of group membership and that is related to outcome is omitted from the selection equation. This omission causes a correlation between the error terms and predictors in the selection and effect estimation equations, which will cause biased estimation of effects. As with regression discontinuity, the functional form of the selection equation must be correctly specified; for instance, if nonlinear or interaction terms that affect group membership are omitted, the effect estimation equation may yield biased estimates.

Selection bias models have been widely studied, praised, and criticized.[8] On the positive side, they address the very important question of taking hidden bias into account rather than just adjusting for observed covariates. And some empirical data can be interpreted positively (Heckman & Hotz, 1989a; Heckman, Hotz, & Dabos, 1987; Heckman & Todd, 1996; Reynolds & Temple, 1995), encouraging further work to develop these models (Moffitt, 1989). Less optimistically, sensitivity to violations of assumptions seems high, and many statisticians are skeptical about these models (e.g., Holland, 1989; Little, 1985; Wainer, 1986). Further, some studies suggest that these models do not well-approximate results from randomized experiments. For example, LaLonde and Maynard (1987) compared results from a randomized experiment with results from a selection bias analysis of the same data using a quasi-experimental control and found that the two answers did not match well. The presumption is that the randomized experiment is correct. Related studies have not yielded promising results (Fraker & Maynard, 1986, 1987; Friedlander & Robins, 1995; LaLonde, 1986; Murnane, Newstead, & Olsen, 1985; Stolzenberg & Relles, 1990; Virdin, 1993).[9] Thus even some economists have been led to prefer randomized experiments to nonrandomized experiments that use selection bias models (Ashenfelter & Card, 1985; Barnow, 1987; Burtless, 1995; Hollister & Hill, 1995). Advocates respond that some of these studies used data that did not meet certain tests for proper application of the models. For example, Heckman and Hotz (1989a, 1989b; Heckman, Hotz, & Dabos, 1987) suggest that a valid selection bias model should find no pretest difference between participants and controls and no posttest difference between randomized and nonrandomized controls (of course, if one has randomized controls, the selection bias estimate is of less interest). But even when those tests are passed, concern about the accuracy of resulting estimates can remain (Friedlander & Robins, 1995).

Work to develop better selection bias models continues (Heckman & Roselius, 1994, 1995; Heckman & Todd, 1996). Bell et al. (1995) note that several events in the 1970s encouraged use of external control groups, like those drawn from national survey archives in selection bias models, and discouraged use of internal controls. Today, there is renewed interest in internal control groups, on the assumption that they may be more similar to the treatment group a priori than are external

8. For a novel view of this debate from sociology of science, see Breslau (1997).

9. Dehejia and Wahba (1999) reanalyzed the Lalonde (1986) data using propensity score analysis and obtained point estimates that were much closer to those from the benchmark randomized experiment.

controls. This point has been widely suggested in the quasi-experimental literature for decades (e.g., Campbell & Stanley, 1963) but not appreciated in selection bias modeling until recently (e.g., Heckman & Roselius, 1994; Heckman et al., 1997). Friedlander and Robins (1995), for example, found that selection bias models of welfare experiments more accurately approximated estimates from randomized experiments when the nonrandomized controls were selected from within the same state as the program recipients rather than from other states. Bell et al. (1995) investigate various internal control groups formed from program applicants who withdrew, were screened out,[10] or did not show for treatment, with encouraging results.

Also, these models would probably work better if they used predictors that were selected to reflect theory and research about variables that affect selection into treatment, a procedure that requires studying the nature of selection bias as a phenomenon in its own right (e.g., Anderman et al., 1995). For example, Reynolds and Temple (1995) obtained effect size estimates from selection bias models that closely resembled those from randomized experiments on the effects of participation in a preschool program. The rules for eligibility to participate in the program were fairly clear, and the authors were able to predict participation fairly accurately. However, when less is known about participation, even authors who have made extensive efforts to select comparable controls and measure pertinent selection predictors have found results that made them question if the selection bias model worked (Grossman & Tierney, 1993).

Heckman (Heckman, Lalonde, & Smith, 1999; Heckman & Roselius, 1994, 1995; Heckman & Todd, 1996) has incorporated these lessons into various revised models that test the effects of employment and training programs under the Job Training Partnership Act (JTPA). The context is the National JTPA Experiment, commissioned in 1986 by the U.S. Department of Labor. It included program participants, a randomized control group, and a nonrandomized comparison group of people who were eligible for JTPA but did apply. Heckman tested several semiparametric selection bias estimators that do not require such strong assumptions, and they performed better than previous parametric models had.[11] However, making fewer assumptions usually results in weaker inferences, and the hard question of which assumptions are appropriate still remains to be solved in each study. In any case, the best performing models in Heckman and Todd used matching on a modified version of the propensity scores described in the previous

10. Bell et al. (1995) refer to this group as a regression discontinuity control group. Close examination of the procedures used to create this control make that unlikely, for two reasons. First, assignment does not seem to have been made *solely* on the basis of a quantitative cutoff. Second, staff who were making program selection decisions may have created the participant's score on the selection variable as a *result* of their judgments about who should get treatment rather than by first measuring the variable and then determining eligibility. Thus both the score and the cutoff may have been the effect of assignment rather than its cause.

11. All of these tests were conducted with knowledge of the outcome of the randomized experiment, so they leave doubt as to how well they would have performed under conditions in which the researcher does not know the correct answer—which is, after all, the likely context of application.

section. Heckman and Todd (1996) note that these matching methods "perform best when (1) comparison group members come from the same local labor markets as participants, (2) they answer the same survey questionnaires, and (3) when data on key determinants of program participation is (sic) available" (p. 60). Perhaps such results indicate the beginnings of a convergence among the statistical, econometric, and quasi-experimental design literatures in understanding of how to get better effect estimates from quasi-experiments.

Another indicator of convergence comes from Manski and his colleagues (e.g., Manski, 1990; Manski & Nagin, 1998; Manski, Sandefur, McLanahan, & Powers, 1992), who have explored nonparametric methods for placing bounds on treatment effects under conditions of selection bias, similar to the sensitivity analysis tradition. These methods do not make the strict assumptions of the parametric methods of Heckman. They result in a series of treatment estimate bounds that vary depending on the assumptions made. But the estimates that require the fewest assumptions also sacrifice power, so the bounds may be unacceptably wide; and point estimates of treatment effects can be attained only by making stronger assumptions about the process generating treatment assignment and outcome, that is, if one or more plausible selection models can be identified. Rosenbaum (1995b) suggests that Manski's bounds are analogous to the limit of a sensitivity analysis in which the key index of potential bias in the sensitivity analysis (Γ) approaches ∞; and he agrees that the bounds are conservative but that they contain some information. Copas and Li (1997) discuss the relationship between selection models and sensitivity analyses, arguing that selection models are so sensitive to assumptions that they should be used as sensitivity analyses by varying those assumptions deliberately rather than being used to estimate single treatment parameters—a point of view with which Heckman and others have expressed sympathy (e.g., Heckman & Hotz, 1986; Winship & Mare, 1992). All agree: sensitivity analyses are crucial in nonrandomized experiments.

Latent Variable Structural Equation Modeling

The work of Karl Joreskog and his colleagues on structural equation modeling (e.g., Joreskog & Sorbom, 1988, 1993) and the similar but more user-friendly work by Peter Bentler (e.g., Bentler, 1993, 1995) have led to widespread use of so-called "causal modeling" techniques. When these techniques were applied to data from quasi-experiments, the hope was to make causal inferences more accurate by adjusting for predictors of outcome that might be correlated with receipt of treatment and by adjusting for unreliability of measurement in predictors. If these two goals could be accomplished, an unbiased estimate of treatment effect could be obtained. In fact, adjustment for measurement error is feasible using latent variable models (see Chapter 12). Doing so requires using multiple observed measures of a construct that are, in essence, factor analyzed to yield latent variables shorn of random measurement error (multiple measurement can take place on a subsample to

save costs; Allison & Hauser, 1991). Those latent variables can be used to model treatment outcome and may improve estimation of treatment effects. For instance, several reanalyses of the original Head Start data (Cicirelli & Associates, 1969) have been done with latent variables, all resulting in estimates of effects that are thought to be better than those the original analysis yielded (Bentler & Woodward, 1978; Magidson, 1977, 1978, 2000; Reynolds & Temple, 1995; Rindskopf, 1981).

However, the other goal of these models, to adjust outcome for variables correlated with treatment assignment and outcome, is problematic because it is rare in the social sciences to have sufficient knowledge of all the treatment-correlated predictors of outcome. The almost inevitable failure to include some such predictors leaves hidden bias and inaccurate effect estimates. Moreover, it is not enough that the model include all the predictors; it must correctly specify their relationship to each other and to outcome, including nonlinear and interaction terms, correct modeling of mediated versus direct relationships, and correct sequencing of lagged relationships. Reichardt and Gollob (1986) provide a readable introduction to the issues; Bollen (1989) has a detailed one; and Bentler and Chou (1988) give practical tips for the more productive use of these models. Ultimately, all agree that these causal models are only as good as the design underlying the data that go into them—even the developers of LISREL, who make clear that their program estimates causal parameters presumed to be true as opposed to testing whether the relationships themselves are causal (Joreskog & Sorbom, 1990).

The literature on structural equation modeling has developed largely independently of the literatures on selection bias modeling and propensity scores. In part, this is an accident of the different disciplines in which these developments first occurred; and in part it is because the methods attempt different adjustments, with structural equation models adjusting for predictors of outcome but with selection bias models and propensity scores adjusting for predictors of treatment selection. Yet Winship and Morgan (1999) make clear that there is a close relationship among all these methods (see also Pearl, 2000; Spirtes, Glymour, & Scheines, 2000). It is unclear whether such efforts at integrating these topics will be successful without making the same kinds of assumptions that have so far stymied prior analyses. But it is also clear that the attention being paid to causal inference is increasing across a wide array of disciplines, which at a minimum bodes well for further integration of this disparate literature.