4

Quasi-Experimental Designs That Either Lack a Control Group or Lack Pretest Observations on the Outcome

> **Qua·si** (kwā′zī′, sī′, kwä′zī, sī′): [Middle English as if, from Old French from Latin *quasi: quam,* as; see *kwo-* in Indo-European Roots + *s,* if; see *swo-* in Indo-European Roots.] adj. Having a likeness to something; resembling: *a quasi success.*

WE BEGIN this chapter (and subsequent ones) with a brief example that illustrates the kind of design being discussed. In 1966, the Canadian Province of Ontario began a program to screen and treat infants born with phenylketonuria (PKU) in order to prevent PKU-based retardation. An evaluation found that, after the program was completed, 44 infants born with PKU experienced no retardation, whereas only 3 infants showed evidence of retardation—and of these three, two had been missed by the screening program (Webb et al., 1973). Statistics from prior years showed a higher rate of retardation due to PKU. Although the methodology in this study was quite primitive, particularly because it lacked a control group[1] that did not receive treatment, the authors concluded that the program successfully prevented PKU-based retardation. Subsequently, such programs were widely adopted in Canada and the United States and are still seen as extremely effective. How did this study achieve such a clear, correct, and useful conclusion when it used neither a control group nor random assignment? That is the topic of this chapter.

[1] The term *control group* usually refers to a group that does not receive treatment; the more general term *comparison group* may include both control groups and alternative treatment groups.

In this chapter, we describe quasi-experimental designs that either lack a control group or that lack pretest observations on the outcome. Although few such studies achieve causal conclusions as clear as this PKU example, it is sometimes possible. More importantly, researchers often have good reasons for using such designs, such as a need to devote more resources to construct validity or external validity; practical necessities imposed by funding, ethics, or administrators; or logistical constraints that occur when an intervention has already been fielded before the evaluation of that intervention is designed. Indeed, given such contingencies, sometimes one of these designs will be the *best* design for a given study, even if the causal inference itself may be weaker than might otherwise be possible. Consequently, in this chapter we present such designs and the conditions that make them more likely to be successful for descriptive causal inference. But we use this chapter to make three more points. First, these designs are used quite frequently in field research; for example, in a recent review of the Even Start Literacy Program (St. Pierre, Ricciuti & Creps, 1998), the vast majority (76%) of studies used a one group pretest-posttest design, and most of the rest used the same design without a pretest. Sometimes such uses reflect a mistaken belief that design elements such as control groups or pretests are undesirable or unnecessary—even when descriptive causal inference is the highest priority. We want to cast doubt on such beliefs by showing the costs to internal validity that such designs can incur so that researchers can choose whether to incur these costs given their other research priorities. Second, we use these designs to illustrate how various validity threats operate with actual examples, for it is more important to learn how to think critically about these threats than it is to learn a list of designs. Successive exposure to these threats here and in later chapters will make it easier to detect them in studies one reads or designs. Finally, we use these designs to introduce the structural elements common to all experimental designs, out of which researchers build designs that are stronger for internal validity to suit the circumstances of their work, elements that are used again and again in the designs we describe in later chapters.

THE LOGIC OF QUASI-EXPERIMENTATION IN BRIEF

The designs in this chapter are **quasi-experiments**—experiments that lack random assignment of units to conditions but that otherwise have similar purposes and structural attributes to randomized experiments. Quasi-experiments have been used for many years. Lind (1753) described a quasi-experimental comparison of six medical treatments for scurvy, and Galton (1872) described a quasi-experimental thought experiment he never actually implemented on the effects of prayer. Hartmann (1936) studied the effects of emotionally versus rationally based political leaflets on election results in Pennsylvania. He matched three voting wards that received the emotional leaflets with four others that received the rational leaflets, matching on size of ward, density of population, assessed real-estate values, previous voting patterns, and socioeconomic status.

A causal inference from any quasi-experiment must meet the basic requirements for all causal relationships: that cause precede effect, that cause covary with effect, and that alternative explanations for the causal relationship are implausible. Both randomized and quasi-experiments manipulate the treatment to force it to occur before the effect. Assessing covariation between cause and effect is easily accomplished in all experiments, usually during statistical analysis. To meet the third requirement, randomized experiments make alternative explanations implausible by ensuring that they are randomly distributed over the experimental conditions. Because quasi-experiments do not use random assignment, they rely on other principles to show that alternative explanations are implausible. We emphasize three closely related principles to address this requirement in quasi-experimentation.

- The first principle is the *identification and study of plausible threats to internal validity.* Once identified, those threats can be studied to probe how likely it is that they explain treatment–outcome covariation (Reichardt, 2000). This chapter provides numerous examples of how to criticize inferences from quasi-experiments using the threats to validity presented in the previous chapters.
- A second principle is the *primacy of control by design.* By adding design elements (e.g., observation at more pretest time points, additional control groups), quasi-experimentation aims either to prevent the confounding of a threat to validity with treatment effects or to provide evidence about the plausibility of those threats. The usual alternative to design controls are statistical controls that attempt to remove confounds from effect estimates using statistical adjustments after the study is done. Design controls and statistics can and should be used together, of course. But we encourage as much prior use of design controls as possible, leaving statistical control to deal with any smaller differences that remain after design controls have been used.
- A third principle for reducing the plausibility of alternative causal explanations in quasi-experiments is *coherent pattern matching.* That is, a complex prediction is made about a given causal hypothesis that few alternative explanations can match. Examples in this chapter include the use of **nonequivalent dependent variables** and of predicted interactions. The more complex the pattern that is successfully predicted, the less likely it is that alternative explanations could generate the same pattern, and so the more likely it is that the treatment had a real effect.

None of these three principles provide the ease of causal inference or the elegant statistical rationale associated with random assignment. Instead, the logic of causal inference in quasi-experimentation requires careful and detailed attention to identifying and reducing the plausibility of alternative causal explanations.

TABLE 4.1 Quasi-Experimental Designs Without Control Groups

The One-Group Posttest-Only Design

$$X \qquad O_1$$

The One-Group Posttest-Only Design With Multiple Substantive Posttests

$$X_1 \qquad \{O_{1A}\ O_{1B} \ldots O_{1N}\}$$

The One-Group Pretest–Posttest Design

$$O_1 \qquad X \qquad O_2$$

The One-Group Pretest–Posttest Design Using a Double Pretest

$$O_1 \qquad\qquad O_2 \qquad X \qquad O_3$$

The One-Group Pretest–Posttest Design Using a Nonequivalent Dependent Variable

$$\{O_{1A}, O_{1B}\} \quad X \qquad \{O_{2A}, O_{2B}\}$$

The Removed-Treatment Design

$$O_1 \qquad X \qquad O_2 \qquad\qquad O_3 \qquad \cancel{X} \qquad O_4$$

The Repeated-Treatment Design

$$O_1 \qquad X \qquad O_2 \qquad \cancel{X} \qquad O_3 \qquad X \qquad O_4$$

DESIGNS WITHOUT CONTROL GROUPS

In this section, we discuss designs without a control group (Table 4.1). Designs without control groups can yield strong causal inferences only by reducing the plausibility of alternative explanations for the treatment effect. Some designs do a poor job of that, and others do better. Progressing from the former to the latter demonstrates how to build designs that render more threats to internal validity implausible.

The One-Group Posttest-Only Design

This design obtains one posttest observation on respondents who experienced a treatment, but there are neither control groups nor pretests. This design is diagrammed as:

$$X \qquad O_1$$

where X is the treatment, O_1 is the posttest, and position from left to right indicates temporal order. The absence of a pretest makes it difficult to know if a change has occurred, and the absence of a no-treatment control group makes it difficult to know what would have happened without treatment. Nearly all the threats to internal validity except ambiguity about temporal precedence usually apply to this design. For example, a history threat is nearly always present because other events might have occurred at the same time as treatment to produce the observed effect.

However, the design has merit in rare cases in which much specific background knowledge exists about how the dependent variable behaves. For example, background knowledge of calculus is very low and stable in the average high school population in the United States. So if students pass a calculus test at levels substantially above chance after taking a calculus course, this effect is likely to be due to the course. Students simply do not learn much calculus from their homes, friends, television sets, recreational activities, or even other academic courses. But for valid descriptive causal inferences to result, the effect must be large enough to stand out clearly, and either the possible alternative causes must be known and be clearly implausible or there should be no known alternatives that could operate in the study context (Campbell, 1975). These conditions are rarely met in the social sciences, and so this design is rarely useful in this simple form.

Improving the One-Group Posttest-Only Design Using Multiple Substantive Posttests

The one-group design without pretests can be more interpretable under theory-linked conditions variously called pattern matching (Campbell, 1966a; Trochim, 1985) or coherence (Rosenbaum, 1995a). Consider the analogy of a crime such as murder. Noting that detectives can be successful in discovering the cause of such crimes (the murderer), Scriven (1976) attributes their success to the saliency of the effect (a dead body), to the availability of a pattern of clues that specify the time and manner of death (multiple posttests), and to the ability to link these clues to the modus operandi of criminals (potential alternative explanations) who are known to commit their crimes in distinctive ways that might overlap with the details found at the crime scene. If more than one person has a particular *modus operandi,* then the known suspects can be questioned to probe their alibis (Abelson, 1995, calls this the method of signatures).

Pathologists use this detective-like approach to investigate why someone died (the effect), using evidence from the corpse, the setting, and the time of death (a pattern of data). They identify possible causes of the death by matching that pattern of data to the descriptions in the scientific literature that differentiate one disease from another. Epidemiologists do something similar. To learn how AIDS arrived in the United States, they used available clues (early prevalence in the homosexual community, high prevalence among Africans living in Europe, and the military involvement of Cuba in equatorial Africa) to tentatively trace the disease's origin to a homosexual Air Canada steward who was very sexually active in the United States and who had visited Cuba and had consorted there with soldiers who had served in equatorial Africa, where AIDS was already endemic.

These clues serve as multiple, unique, and substantive posttests to the design:

$$X_1 \qquad \{O_{1A}\ O_{1B} \ldots O_{1N}\}$$

one he wanted to attribute to the cause—he found lower lead levels after treatment, which also happened to be in winter. So those lower lead levels could have been due to seasonality effects. Both these conditions are necessary to weaken the inference that the program caused the observed effects. Imagine, for example, that past research showed that lead dust levels were lower in summer than in winter. This could not possibly explain his finding of lowered levels in winter and so could not threaten internal validity. For a threat to validity to be plausible, it must produce an effect that is similar in size to the one observed, not of a magnitude that is too small to explain a large observed effect and not an effect in the opposite direction (e.g., not an increase in dust when a decrease was observed).

However, social scientists in field settings will rarely be able to construct confident causal knowledge with the simple pretest–posttest design unless the outcomes are particularly well behaved and the interval between pretest and posttest is short. Persons considering this design should consider adding even more design elements.

Improving the One-Group Pretest–Posttest Design Using a Double Pretest

The plausibility of maturation and regression threats is reduced by adding a second pretest prior to the first pretest:

$$O_1 \qquad O_2 \quad X \quad O_3$$

The two pretests function as a "dry run" to clarify the biases that might exist in estimating the effects of treatment from O_2 to O_3. For example, Marin, Marin, Perez-Stable, Sabogal, and Ostero-Sabogal (1990) tested the effects of a culturally appropriate smoking cessation information campaign for Hispanics. The first pretest occurred in the fall of 1986, the second pretest in the summer of 1987, and the posttest in the summer of 1988. Results showed that information levels after the campaign (O_3) were far higher than before it (O_2) and also far higher than the maturational trend from O_1 to O_2 would yield. Of course, a nonlinear change in maturation trend before treatment could be detected only with even more pretests.

Improving the One-Group Pretest–Posttest Design Using a Nonequivalent Dependent Variable

The addition of a nonequivalent dependent variable is diagrammed below, where *A* and *B* represent different measures collected from a single group at times 1 and 2:

$$\{O_{1A}, O_{1B}\} \quad X \quad \{O_{2A}, O_{2B}\}$$

Measures *A* and *B* assess similar constructs. Measure *A* (the outcome) is expected to change because of treatment. Measure *B* (the nonequivalent dependent variable) is not; however, Measure *B* is expected to respond to salient internal valid-where $\{O_{1A}\ O_{1B}...O_{1N}\}$ refers to measures of different posttreatment constructs A through N that are matched to the pattern of effects (the modus operandi) left by the possible causes that are known (i.e., the suspects). This is very different from a design in which only one posttest construct (O_{1A}) is assessed so that a pattern-matching logic cannot be used.

However, in all these examples, the effect is known, but the cause is unknown and is searched for retrospectively. In most quasi-experimentation, the situation is the opposite: the potential cause is known (e.g., a new calculus course) but the effect is unknown and is searched for prospectively (e.g., what happens to student achievement). In this latter case, the pattern-matching logic is less compelling because the cause is often an innovation with an unknown pattern of effects. So adding multiple posttests under a prospective design can increase Type I errors, doubly risky because humans are adept at finding and interpreting patterns even in random data (Fischhoff, 1975; Paulos, 1988; Wood, 1978). Careful *prior* specification of patterns that support a causal relationship is crucial and sometimes occurs in fields with well-developed theory. But even then it is crucial that the predicted pattern be unique. If pathology textbooks show similar changes associated with three different diseases, we cannot discriminate among those diseases based on those changes.

The One-Group Pretest–Posttest Design

Adding a pretest measure of the outcome construct to the preceding design yields a one-group pretest–posttest design. A single pretest observation is taken on a group of respondents (O_1), treatment (X) then occurs, and a single posttest observation on the same measure (O_2) follows:

$$O_1 \quad X \quad O_2$$

Adding the pretest provides weak information about the counterfactual inference concerning what might have happened to participants had the treatment not occurred. However, because O_1 occurs before O_2, the two may differ for reasons unrelated to treatment, such as maturation or history. For example, Jason, McCoy, Blanco, and Zolik (1981) studied the effects of a campaign to reduce dog litter in Chicago by distributing to residents both educational material and "pooper-scoopers" with plastic bags. The subsequent reduction in dog litter was dramatic. The outcome can be causally tied to the intervention to the extent to which one can assume that other possible alternative hypotheses are implausible—for example, that the weather had turned worse and dogs were on the street less often or that a citywide campaign to reduce litter had begun at the same time or that there had been a dramatic increase in local crimes that kept residents indoors.

The design can be implemented either with the same units or different units receiving both pretest and posttest. When the same units are used, this is often called a within-participants design.[2] Duckart (1998) used the within-participants version of this design to evaluate the effects of a program to reduce environmental lead in low-income urban housing units in Baltimore. The program used both education and physical changes to reduce sources of lead in each home. Using lead wipe samples, lead levels were measured at various locations in each home at pretest, immediately after the intervention (posttest), and 6 months later (follow-up). Significant reductions occurred in lead levels between pretest and posttest and to a lesser extent at follow-up. Because lead levels are stable over short periods of time, spontaneous changes that a control group might detect are unlikely to occur, so this topic lends itself to this design better than most other topics.

Even so, Duckart (1998) identifies several threats to validity for this design. Regarding internal validity, maturation may affect changes from pretest to follow-up because lead dust levels are lower in winter than in summer; and most of the follow-ups were conducted in winter. History is a threat because another Baltimore agency concurrently provided services that could have affected lead levels at some homes in the sample. Testing is a threat because pretest feedback to residents about lead levels may have caused them to clean more diligently, which would reduce lead levels even if the intervention were ineffective. Regarding attrition, about one third of participants did not continue beyond pretest, and it may have been the least motivated or cooperative residents who dropped out. Statistical conclusion validity may have been reduced by low sample sizes and consequent low power for those tests that were not significant, and power could have been further reduced because treatment **implementation** varied considerably from site to site. Construct validity of the outcome measures may have been hampered because the same persons who administered the intervention in the homes also administered the outcome measures, and they may have done so in a way that inadvertently made the results look favorable to the intervention.

This example is particularly instructive about the logic of applying threats to the validity of a causal inference. Duckart (1998) did two key things. First, he showed that the threats were more than just possible but were plausible. He did this by providing data to show that a threat was likely to have occurred in his study—past research shows that lead dust levels are indeed lower in winter than in summer. Second, he showed that this threat would produce an effect like the

[2] Here the within-participants factor is time: repeated measures at pretest and posttest within each unit. In other within-participant designs, more than one condition can be administered to the same units. By contrast, an experiment having multiple conditions and different units in each condition is often called a **between-participants design** (S. E. Maxwell & Delaney, 1990). Within-participants designs can increase statistical power by controlling individual differences between units within conditions, and so they can use fewer units to test the same number of treatments. However, within-participants designs can cause **fatigue effects, practice effects, carryover effects,** and **order effects.** To avoid confounding such effects with treatment, the order of treatments in a within-participants design is often either randomized for each unit or deliberately **counterbalanced** so that some units get treatments in one order (e.g., A then B) but others get a second order (B then A) so that order effects can be assessed.

ity threats in the same way as Measure *A*.[3] For example, a nonequivalent dependent variable was used by Robertson and Rossiter (1976) to study children's preferences for advertised toys during the Christmas marketing period. They showed that toys advertised in November and December increased in desirability (the outcome) more than nonadvertised toys (the nonequivalent dependent variable). This design reduces the plausibility of many threats to internal validity. For example, history is a threat, given the frequent references to gifts and toys at that time of year in American culture. However, this should affect preferences for all toys, not just advertised toys. A different history problem arises, though, if toys advertised on television are supported by ads on radio and in newspapers. Who can say whether television advertising caused the effect? Or what about statistical regression if toy manufacturers advertised those toys that were selling poorly during this period?

McKillip and Baldwin (1990) also used a nonequivalent dependent variable design and found that awareness of condom use increased more than awareness of alcohol abuse or of regular exercise after a media campaign about sexually transmitted diseases; and McNees, Gilliam, Schnelle, and Risley (1979) found that feedback about theft of potato chips from a snack bar reduced such theft but not theft of ice cream, milk, or sandwiches. This design is broadly useful and can often be causally interpretable when both dependent variables are plausibly exposed to the same set of environmental causes to the same degree.

The Removed-Treatment Design

This design adds a third posttest (O_3) to the one-group pretest–posttest design (O_1 and O_2) and then removes the treatment ($\cancel{X}$ symbolizes this removal) before a final measure is made (O_4):

$$O_1 \quad X \quad O_2 \qquad O_3 \quad \cancel{X} \quad O_4$$

The aim is to demonstrate that outcome rises and falls with the presence or absence of treatment, a result that could be otherwise explained only by a threat to validity that similarly rose and fell over the same time. The link from O_1 to O_2 is one experimental sequence, and the change from O_3 to O_4 is another that involves the opposite hypothesis to the O_1 to O_2 hypothesis. If the first sequence predicts, say, an increase from O_1 to O_2, the second sequence predicts a decrease (or a smaller increase). Of course, this pattern can occur only if the treatment effects

[3] Rosenbaum (1995a, p. 137) references a similar idea when he notes "a systematic difference between treated and control groups in an outcome the treatment does not affect must be a hidden bias"; however, nonequivalent dependent variables are chosen to respond to particular threats to validity, providing more information about specific alternative causes than other outcomes that the treatment should not affect.

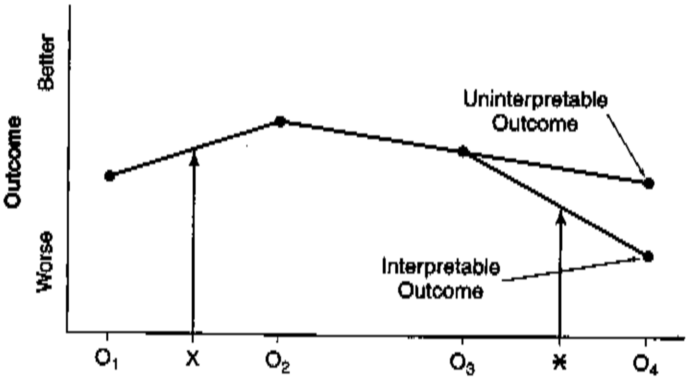

FIGURE 4.1 Generally interpretable outcome of the removed-treatment design

dissipate once treatment is removed. Even a partially persisting effect from O_2 to O_3 biases the analysis against the reversal of the treatment effect between O_3 and O_4. Thus the most interpretable outcome of the design is presented in Figure 4.1.

Statistical conclusion validity is a problem, because the pattern of results can be influenced by even a single outlier. So large sample sizes and reliable measures are a desideratum. Further, removing some treatments might be unethical or might arouse frustration that would be correlated with measures of aggression, satisfaction, or performance, operating similarly to resentful demoralization and compensatory rivalry. If so, it is unwise to use this design.

A naturally occurring version of this design occurs when respondents stop taking treatment, but only if their reasons for discontinuing are unrelated to treatment. That is rarely the case, so special care must be taken when respondents self-select out of treatment. To envision this, consider a study of the effects on attitudes of entering a new job role. For example, someone becomes a foreman (*X*); he or she develops promanagerial attitudes between O_1 and O_2 but dislikes the new contact with managers. By O_3 he or she has become less promanagerial (Lieberman, 1956). Such a person could then resign from the foremanship or be relieved of it, leading to less promanagerial attitudes from O_3 to O_4 when compared with O_1 to O_2. Given this pattern, the researcher must decide whether the O_3 to O_4 decrease reflects the job change or was already occurring by O_3. The latter would be more likely if the O_3 to O_4 difference were similar in size to the O_2 to O_3 difference (see the uninterpretable outcome in Figure 4.1). However, a steeper decrease in promanagerial attitudes between O_3 and O_4 when compared with O_2 to O_3 would suggest that entering the new role causes one to adopt promanagerial attitudes.

Lieberman (1956) actually used a simpler version of this design. He sampled men before they became foremen, after they became foremen, and then after they reverted to worker status—only three measurement waves. Hence, differences between O_1 and O_2 and O_2 and O_3 might be due to (1) a role change that influenced attitudes or (2) the demotion of foremen whose attitudes were becoming less managerial. Adding the fourth observation helps assess these possibilities.

Making observations at equally spaced intervals is important with this (and many other) designs because it permits assessment of spontaneous linear changes over time. Comparison of the differences between O_2 and O_3 and between O_3 and O_4 would be less meaningful if the O_3 to O_4 time interval were longer than the O_2 to O_3 interval because a constant rate of change would reveal larger O_3 to O_4 differences than O_2 to O_3 differences. If one has a *treatment-free* estimate of rate of change per time interval, the need for equal spacing is less; but an estimate of that rate is rarely possible with this design.

The Repeated-Treatment Design

It is sometimes possible to introduce, remove, and reintroduce treatment over time to study how treatment and outcome covary over time:

$$O_1 \quad X \quad O_2 \quad \cancel{X} \quad O_3 \quad X \quad O_4$$

Even more so than in the previous design, few threats to validity could explain a close relationship between treatment introductions and removals on the one hand and parallel changes in outcome on the other. Such threats would have to come and go on the same schedule as treatment introduction and removal. The latter is usually an unlikely event. However, if treatment effects are not transient, it is difficult for treatment removal to reverse the direction of effect; and if treatment creates a ceiling effect by O_3, that would prevent the reintroduction of the treatment from registering an effect.

The most interpretable outcome of this design is the case in which O_1 differs from O_2, O_2 differs from O_3 in the opposite direction, and the O_3 to O_4 difference resembles the O_1 to O_2 difference, not the O_2 to O_3 difference. Powers and Anglin (1993) adapted this design to demonstrate the effects of methadone maintenance on narcotic use. Narcotic use dropped dramatically while methadone was administered, rose when methadone was removed, and dropped again when methadone was readministered. This design is also frequently used by behavioral researchers in psychology (Barlow & Hersen, 1984). We suspect that this design has been used so much because it involves at least one replication of the treatment effect and because this replication meets a basic criterion for quality research: reproducibility.

A threat to internal validity is cyclical maturation. For example, if O_2 and O_4 were recorded on Tuesday morning and O_1 and O_3 on Friday afternoon, differences in productivity might be related to day-of-the-week differences in performance rather than to treatment. The researcher should also check whether unique historical events follow the pattern of treatment introduction and removal. In general, though, the design is strong for internal validity, especially when the investigator controls introduction and removal of treatment.

The design can be vulnerable on external and statistical conclusion validity grounds. For example, many performance graphs in the Hawthorne studies (Roethlisberger & Dickson, 1939) are of individual female workers, sometimes as few as six women, who displayed considerable variability in how they reacted to the treatments; so we cannot be sure how robust the results are to sampling error. Of course, this design can use larger samples and statistical tests, which we encourage.

Construct validity of the cause is threatened when respondents notice the introduction, removal, or reintroduction of the treatment. Respondents may generate hypotheses about these treatment variations and respond to them. Resentful demoralization can also be an issue when the treatment is removed between O_2 and O_3. The O_3 data point might then be affected, making it more difficult to interpret an increase between O_3 and O_4 when the treatment is reinstated.

So this design is better with transient effects, with unobtrusive treatments, with a long delay between initial treatment and its reintroduction, and with no confounding of temporal cycles with the treatment's introduction, removal, and reintroduction. It is also more effective when reintroductions of the treatment are frequent and randomly distributed across time, thus creating a randomized experiment in which time blocks are the unit of assignment (Edgington, 1987, 1992). Not all research projects can meet these conditions.

Implementing the Repeated-Treatment Design Using Retrospective Pretests

Powers and Anglin (1993) actually adapted this design by assessing retrospectively the effects of methadone maintenance on heroin addicts who had experienced multiple treatment episodes in the past. They asked addicts to recreate their drug use and treatment history retrospectively. They found through these recollected histories that methadone maintenance reduced drug use during treatment periods but not between them. Unfortunately, retrospective judgments can be quite biased (e.g., Silka, 1989). For example, people tend to overestimate "bad" occurrences in the present compared with the past, and results from retrospective self-reports can differ from prospective self-reports (Widon, Weiler, & Cottler, 1999; G. S. Howard et al., 1979; G. S. Howard, Millham, Slaten, & O'Donnell, 1981). Factors that influence the retrospective pretest include whether the material is easily distorted (e.g., cognition versus behaviors), length of time since the events being recalled, demand characteristics (e.g., distorted responses of illegal behaviors), specificity versus generality of information needed (specific events are less accurately recalled), and the emotions elicited by the recall (e.g., remembering a trauma; Babcock, 1998). Sometimes retrospective pretests can be cross-validated by other sources. For example, Powers and Anglin (1993) validated self-reported treatment dates using stored administrative records. However, little is known empirically about these matters in the context of experimental design, so we repeat the opinion of Campbell and Stanley (1963): "Given the autistic factors known to distort memory and interview reports, such data (i.e., retrospective pretest data) can never be crucial" (p. 66). Retrospective pretests should be a supplement to other design improvements, not used by themselves, and should be a method of last resort interpreted with great caution.

The Powers and Anglin (1993) study also had a problem of attrition over time. They could study only addicts who returned for treatment. Thus their sample sizes of addicts with one, two, three, or four treatment episodes were progressively smaller. If some patients returned to treatment because they had failed to stay clean, whereas others did not return because they had successfully withdrawn from heroin, then the general conclusion that methadone maintenance results in a success that is only temporary might not have held if the full sample had been followed. This is a general problem in any longitudinal study that follows the same people over time. Those who are not reinterviewed over time may differ systematically from those who are reinterviewed.

Many of the problems we have discussed in this section could be addressed better if an independent control group had been included. In the Powers and Anglin (1993) study, for example, the inclusion of addicts who were not on methadone maintenance might have clarified the frequency and timing of changes in retrospective pretest reports in untreated clients and might have clarified the effects of attrition over time. In the latter case, two controls could have been used: one that was required to return to the clinic for assessment in which the same sort of attrition might occur as in treated clients and one that was followed aggressively in the community, where such attrition might not occur. We turn now to discussion of such control groups, limiting discussion this time to quasi-experiments that have control groups but do not have pretests.

DESIGNS THAT USE A CONTROL GROUP BUT NO PRETEST

A classic method for supporting a counterfactual inference is to add a control group that receives no treatment, with the control group selected to be as similar as possible to the treatment group (D'Agostino & Kwan, 1995). Notationally, a dashed line (---------) between groups indicates that they were not randomly formed, and such groups are preceded by *NR* (nonrandom assignment). Table 4.2 summarizes quasi-experimental designs that have control groups but no pretest measures on the outcome variable.

Posttest-Only Design With Nonequivalent Groups

Here we add a control group to the one-group posttest-only design. This design might be used if a treatment begins before the researcher is consulted so that

TABLE 4.2 Quasi-Experimental Designs That Use Control Groups But No Pretest

Posttest-Only Design With Nonequivalent Groups			
NR		*X*	O_1
NR			O_2
Posttest-Only Design Using an Independent Pretest Sample			
NR	O_1	*X*	O_2
NR	O_1		O_2
Posttest-Only Design Using Proxy Pretests			
NR	O_{A1}	*X*	O_{B2}
NR	O_{A1}		O_{B2}

pretest observations are not available on the scale used at posttest. Such a design can be diagrammed as:

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

For example, Sharpe and Wetherbee (1980) compared mothers who received nutritional benefits from a Women, Infants, and Children (WIC) project in Mississippi with those who did not. Results indicated no significant posttest differences in infant birth weight or infant mortality between the two groups. However, groups may have differed on many variables related to these outcomes, such as prior nutrition status. This possibility of pretest group differences makes it very hard to separate treatment from selection effects.

A rationale sometimes offered for preferring this weak design is that pretest measurement may sensitize participants and so influence their posttest scores (Lana, 1969). But when different treatment groups are being compared, testing effects that are constant across groups do not threaten internal validity. Only differential testing effects do. They are rare, though they do occur (Aiken & West, 1990). They can be reduced by administering alternate forms of a test (one at pretest and the other at posttest), by using item response theory to calibrate different tests to the same scale (Hambleton, Swaminathan, & Rogers, 1991), by lengthening the time interval between pretest and posttest (Willson & Putnam, 1982), by using a Solomon Four Group Design to assess the presence and impact of such effects,[4] by using unobtrusive measures that are less reactive than self-report (Webb, Campbell,

[4] In this design, participants are randomly assigned to one of two otherwise identical experiments, one with a pretest and one without, so the effect of the pretest can be measured empirically.

Schwartz, & Sechrest, 1966; Webb, Campbell, Schwartz, Sechrest, & Grove, 1981), by using techniques such as the bogus pipeline (Jones & Sigall, 1971), by using retrospective pretests, and by using explicit reference groups or behavioral criteria to anchor responding (Aiken & West, 1990). Thus, even if differential pretest sensitization is a problem, eliminating the pretests can entail much larger costs for detecting selection biases than dealing with them in the ways outlined here.

Improving the Posttest-Only Design Using an Independent Pretest Sample

Even when it is impossible to gather pretest data on the same sample both before and after treatment, one can sometimes gather pretest information from a randomly formed *independent* sample: a group that is drawn randomly from the same population as the posttest sample but that may have overlapping membership.[5] Such groups are useful when pretest measurements may be reactive, when it is too difficult or expensive to follow the same people over time, or when one wishes to study intact communities whose members change over time. The design is diagrammed here, with the vertical line indicating sample independence across time.

$$\begin{array}{ll|ll} NR & O_1 & X & O_2 \\ \hline NR & O_1 & & O_2 \end{array}$$

This design is frequently used in epidemiology, public health, marketing, and political polling and is preferable to a design without pretest data. However, if independent pretest and posttest samples are not randomly sampled from the same population, this design can introduce considerable selection bias into estimates of treatment effects. Nor is selection bias completely avoided by random sampling. First, random selection equates pretest and posttest only within the limits of sampling error, so comparability is more difficult to achieve with small, heterogeneous samples. Second, the populations being sampled may change in qualitative composition between measurement waves, particularly when waves are far apart in time, and those population changes can masquerade as treatment effects. Also, most internal validity threats that we describe in the next chapter as applying to control group designs with *dependent* pretest and posttest samples also apply when *independent* pretest and posttest samples are used. Finally, statistical conclusion validity can be lessened because independent samples at each measurement wave no longer serve as their own within-group statistical controls. Hence this design is recommended only when the need for independent groups is compelling or when problems with dependent groups are severe. If this design is necessary, the researcher should pay special attention to sample size, to how the sampling design is implemented, and to how comparability can be assessed using measures that are stable and reliable (Feldman & McKinlay, 1994).

[5] In some literatures, these are called cross-sectional panels.

Improving the Posttest-Only Design Using Proxy Pretests

An alternative technique is to measure proxies for pretests—variables that are conceptually related to and correlated with the posttest within treatments. The design is diagrammed here, with *A* representing the proxy pretest and *B* the posttest:

$$\begin{array}{cccc} NR & O_{A1} & X & O_{B2} \\ \hdashline NR & O_{A1} & & O_{B2} \end{array}$$

Preferably, proxies should be conceptually related to outcome, not just readily accessible measures such as age, gender, social class, or race. For instance, when evaluating a calculus course for students who have never had calculus, a calculus pretest would yield little pretest variability; a proxy pretest of mathematical aptitude or achievement in algebra would be better. To the extent that proxies are correlated with the posttest, they index how much groups might have differed at pretest in ways that might be correlated with outcome (selection bias), and they index how much dropouts differ from those who stay in the study within and between groups (attrition bias). Even so, the indexing will usually be poorer than a pretest on the outcome variable itself.

Improving the Posttest-Only Design Using Matching or Stratifying

Lack of a pretest causes lack of knowledge about selection biases. Researchers often try to decrease the odds of such biases by forming treatment and control groups through matching or stratifying on likely correlates of the posttest.

Definitions. When **matching,** the researcher groups units with similar scores on the matching variable, so that treatment and control groups each contain units with the same characteristics on the matching variable.[6] For example, Levy, Matthews, Stephenson, Tenney, and Schucker (1985) studied the effects of a nutrition information program on product sales and market share in supermarkets owned by the same chain with a common set of management procedures. They used 10 treatment stores in Washington and 10 control stores in Maryland, creating 10 cross-state pairs, each matched on store size and socioeconomic characteristics because these variables predict sales and market share.

Twin studies provide a very special case of matching (e.g., Ashenfelter & Krueger, 1994). The presumption is that twins are more similar to each other than they are to other people, more so for identical twins, who share the same genetic

[6] Some authors distinguish between blocking and matching, using blocking to indicate groups with similar scores and matching to identify groups with identical scores. In both cases, the number of units in a match-block is the same as the number of conditions in the experiment—for example, matched pairs with two conditions (such as treatment versus control) or matched triplets with three conditions (such as two different treatments and a control). We generally use the terms *block* and *match* interchangeably.

structure, than for fraternal twins, who share only some of the same genetic structure. In addition, twins are usually exposed to common environmental influences, for instance, being raised in the same family by the same parents with the same socioeconomic background. All these genetic and environmental similarities make using twins as matches in a quasi-experiment a very powerful approach.

A closely related technique is **stratifying,** which occurs when units are placed into homogeneous sets that contain more units than the experiment has conditions. An example is stratifying on gender; clearly, it is impossible to obtain a closer match than "male" among a large group of males, so a block will contain many more males than the number of experimental conditions. Sometimes, strata are formed from continuous variables, for example, by dividing achievement test scores at the median to form two large strata. With such a stratification, groups are less homogeneous than they would have been if participants had been matched because participants have more diverse achievement test scores in each stratum. If strata must be used, more strata are usually better than fewer; and five strata are usually sufficient to remove 90% of the variance that would have been accounted for with matching (Cochran, 1968). In the section that follows, we speak mostly of matching, but the same points usually apply to stratifying. Further, some of the methods we describe shortly, such as optimal matching, blur the conceptual distinction between matching and stratifying, though the practical implications should still be obvious.

Methods for Matching. Diverse methods exist for matching (Cochran, 1983; Cochran & Rubin, 1973; Costanza, 1995; Dehejia & Wahba, 1999; Gu & Rosenbaum, 1993; Heckman, Ichimura, & Todd, 1997; Henry & McMillan, 1993; Marsh, 1998; Rosenbaum, 1995a; H. Smith, 1997). Exact matching requires units to have exactly the same score within a match. However, some units will not have an exact match if samples are small, if the distribution of participants between groups on the matching variable is uneven, or if variables are measured using very fine gradations. In caliper matching, the scores need not be identical but must be within a defined distance of each other (Cochran, 1965), though there are different ways of measuring that distance, such as nearest neighbor matching or Mahalanobis distance matching (Hill, Rubin, & Thomas, 2000; Rosenbaum, 1995a).

Sometimes more control units than treatment units exist, so if the researcher can select multiple controls (Henry & McMillan, 1993), it may be possible to improve the match and improve statistical power. For instance, index matching selects multiple control units above and below a treatment unit. Cluster group matching uses cluster analysis to embed the treatment group in a cluster of similar control units. Benchmark group matching selects control units that fall close to the treatment unit on a multivariate distance measure. Simulations suggest that cluster and benchmark methods may work better than index matching (Henry & McMillan, 1993). Finally, in optimal matching, each treatment unit may have multiple matched controls and vice versa (Bergstralh, Kosanke & Jocobsen, 1996; Rosenbaum, 1995a). A definitive review of the strengths and weaknesses of each matching method has yet to be done, with each compared with the other and with alternatives such as the analysis of covariance.[7]

Problems with Matching. Matching has a beleaguered history in quasi-experimentation because there is always the possibility of selection bias.[8] The minimum risk is *undermatching* because some nonredundant predictors of outcome were not included in the matching methodology. For example, although Levy et al. (1985) matched stores on two variables, other variables, such as the stores' product line and proximity to other stores, might have discriminated even more precisely between stores and might have been correlated with outcomes. If additional matching variables had been used, even greater equivalence might have been created between treatment and control stores. Further, because matching can never induce equivalence on variables not used in the matching, additional selection bias can never be ruled out with complete confidence.

However, skepticism about matching is due less to the concern for undermatching (which, after all, still gets one closer to the right answer) than for the possibility that matching may produce a result that is further away from the right answer than if matching had not been used at all. Campbell and Erlebacher (1970) showed how a common form of matching may have had the latter result. Their example was prompted by an evaluation of Head Start by Cicirelli and Associates (1969; Magidson, 2000) that seemed to show that Head Start children ended up doing worse than matched controls. Campbell and Erlebacher showed how this result could be caused by matching if the populations being matched do not overlap completely on the matching variable, so that matches for Head Start are taken from a different end of their distribution (say, the high end) than matched controls, who might be taken from the lower end of their distribution. If that variable is measured with error or is imperfectly correlated with outcome, statistical regression will occur that makes Head Start look harmful.

Marsh (1998) gives a similar example from the evaluation of programs for gifted and talented children in which the latter children come from a higher performing population than available controls. Matches could therefore be obtained from the overlap between the lower end of the gifted and talented population (the end with more negative random error) and the upper end of the control population (the end with more positive error). At posttest, the gifted and talented program children regress upward toward their population mean, and the controls regress down toward theirs. The resulting bias makes gifted and talented programs look effective even if they have no effect at all.

[7] We discuss some relative advantages of matching and the analysis of covariance (ANCOVA) in the appendix to the next chapter and in the chapters on randomized experiments.

[8] This discussion mostly applies to matching in quasi-experiments; in randomized experiments, matching is often a very useful adjunct to random assignment, as we discuss later in this book.

Principles to Guide Better Matching. A main lesson from these examples is that matching in quasi-experimentation works least effectively—and may be more harmful than helpful—when it is done on an unstable or unreliable variable and when the nonequivalent groups from which the matched sets are drawn are increasingly dissimilar when matched. Two methods help counteract these problems. The first is to select groups that are as similar as possible before matching, as much as the context and research question allow. If the distributions of the two groups on the matching variables overlap substantially, then many matches can be obtained without selecting extensively from opposite and extreme tails of the distributions. For example, nonequivalent groups might have more overlapping distributions if the control group is composed of applicants to the treatment group who would have been eligible but who applied too late than they would have if the control group is composed of those who are not eligible at all. When such selection is not possible, examination of the overlap of the two distributions will help alert the researcher to the possibility of regression among the matches.

The second method is to use matching variables that are stable and reliable. Some variables, such as gender and age, are measured with little error, and so they make good matching variables *if* they are correlated with outcome. The reliability of other matching variables can often be improved by aggregation—for example, by creating a composite of many pretest variables taken at the same time (e.g., the propensity score approach described in Chapter 5), by creating a composite of individuals (e.g., using school averages rather than individual student data), and by averaging two or more consecutive pretest performances rather than just one. This latter procedure also helps prevent the researcher from selecting groups *because* their pretest scores are extreme on one occasion (which also might result in statistical regression), given that random errors will tend to cancel out as more and more observations are averaged.

For example, we are much more optimistic about the kind of matching implemented in a recent study by Millsap, Goodson, Chase, and Gamse (1997) of the effects of a school development program (Comer, 1988) on student achievement in Detroit. They used a *stable matched bracketing* methodology in which 12 treatment schools were compared with 24 matched comparison schools. Matching was done (1) on variables that were measured reliably, in this case subdistrict location, school-level achievement test scores, and racial composition (achievement test scores *were* a true pretest in this case, but we ignore this for the sake of illustration); (2) by averaging over several years rather than just one year for the latter two variables; and (3) by using aggregate (e.g., school) rather than individual data. In addition, from a pool of four to six potential matches for each treatment school, two matching comparison schools were selected to bracket each treatment school: one that performed slightly above the treatment school on the prior achievement matching variable and another that fell just below it. Using two comparison schools for each treatment school increased study power at less cost than adding more treatment schools, because the expensive treatment did not have to be implemented in the comparison schools. The increase in power is particularly important when aggregates (e.g., schools) are being studied and when few aggregates are available.

Matching on more than one variable simultaneously becomes more difficult the more variables are used. However, reducing those variables to a multivariate composite makes matching more feasible. The previously described multivariate distance matching (Henry & McMillan, 1993) uses such a composite, as does matching on **propensity scores** (e.g., Rosenbaum 1995a; Dehejia & Wahba, 1999). A propensity score is obtained using logistic regression to predict group membership from pertinent predictors of group membership. Matching on the propensity score minimizes group differences across all the observed variables used in the propensity score equation and is robust over any **functional form** between propensity score and outcome. Propensity score matching can be supplemented by sensitivity analyses designed to probe whether the observed effect would be robust to biases of particular sizes. We discuss propensity scores and **hidden bias** analyses more extensively in an appendix to the next chapter.

The problems that matching must overcome in quasi-experiments are significant. Matching can occur only on observed measures, so hidden bias may remain. Removing unreliability will reduce the likelihood of regression artifacts, but such artifacts can occur when both the matching variables and the outcome are perfectly measured if they are still not perfectly correlated (Campbell & Kenny, 1999). Also, some threats to internal validity occur after pretest, as with history. There can be no direct match for this in the design planning. Still, we are modestly encouraged that the better matching procedures we have reviewed in this section will be more successful than the simplistic approaches to matching based on a single unreliable variable and on different populations that Campbell and Erlebacher (1970) rightly critiqued. Researchers who use matching should take advantage of these better procedures—the days of simple matching on single variables that are not reliably measured should be left completely behind.

Improving the Posttest-Only Design Using Internal Controls

Internal control groups are plausibly drawn from a population similar to that from which the treatment units are taken (Heinsman & Shadish, 1996; Shadish & Ragsdale, 1996). For example, Aiken, West, Schwalm, Carroll, and Hsiung (1998) used both a randomized design and a quasi-experiment to test the effects of a remedial writing program. Their nonequivalent internal control group was composed of eligible students who registered too late to be in the randomized experiment— a group that is plausibly similar to those eligible students who registered on time. In such cases, fewer selection biases are likely to be present than if the control group were external, such as students from another university or those whose ACT scores made them ineligible for the program. Internal controls do not guarantee similarity; for example, in some psychotherapy research the use of treatment acceptors (or completers) as a treatment group and treatment refusers (or dropouts) as a control group suggests obvious selection problems that might influence estimates of treatment effects. So there is no substitute for careful consideration of such potential selection biases when choosing good control groups.

Baker and Rodriguez (1979) used an internal control to study the effects of diverting criminal court defendants from the legal system to social and educational services. Legal aid attorneys objected to random assignment on legal grounds. However, twice as many clients were referred to diversion as could be accepted, so Baker and Rodriguez used a two-step assignment in which (1) time was divided into randomly distributed blocks of 11, 13, 15, 17, 19, or 21 hours in length and (2) the first 50% of clients expected during each block were assigned to treatment, with the remainder assigned to control. Program staff were unaware of the number of hours in the current block, so they could not easily predict when the 50% quota was met or whether the next client would be assigned to treatment. Nor could the courts funnel some clients preferentially to treatment, because they also did not know which block was in effect and because referrals came from multiple courts. Analysis of group differences suggested that treatment and control conditions were similar at pretest on the measured variables. This is a much stronger control than one from a nondiverted population. The random time feature probably increases the randomness of assignment if program staff were unaware when a new time block took effect and if research staff rather than program staff made assignments.

However, these conditions may often have been violated in a way that allowed program staff to selectively funnel certain cases into treatment. After all, the quota was always larger than one. So staff knew that when one referral was assigned to treatment, the next almost certainly would be, too; and assignment can easily be biased when program staff has knowledge of the next assignment (Chalmers, Celano, Sacks, & Smith, 1983; Dunford, 1990). The procedure also requires that the supply of referrals exceed program capacity. This requirement limits the situations in which it can be applied. Still, the procedure warrants use when better options are not feasible. It resembles two designs that we will later see are very strong: (1) a regression discontinuity design in which order of referral is the selection variable and (2) a study with random assignment of time blocks to conditions.

Improving the Posttest-Only Design Using Multiple Control Groups

It is often possible to use *multiple* nonequivalent control groups, as we pointed out in our discussion of matching. For example, Bell, Orr, Blomquist, and Cain (1995) compared a job training intervention group with four comparisons: those who failed to apply for the program, rejected applicants (screenouts), accepted applicants who failed to start treatment (no shows), and applicants who started treatment but dropped out before it ended (dropouts). Using multiple control groups can help in several ways. If the control groups differ from each other as much as they do from the treatment group, these differences obviously could not be caused by the treatment, and so they can index the magnitude of hidden biases that may be present (Rosenbaum, 1995a). If the direction of bias in each control group is known compared with the treatment group, it may be possible to bracket the treatment effect within a range of known biases. For example, Campbell (1969a) discussed both systematic variation controls and bracketing controls. In the former case, the researcher identifies a key threat to validity and then selects multiple controls that span the plausible range of that threat. If observed effects do not vary over that range, the plausibility of the threat is reduced. With bracketing, the researcher selects both a group expected to outperform the treatment group *if the treatment has no effect* and another group expected to underperform it. If the treatment group outperforms both groups, causal inference is strengthened.

Rosenbaum (in press) cites a similar example using multiple controls by Zabin, Hirsch, and Emerson (1989) on the effects on black teenage women of having an abortion. Zabin et al. compared pregnant women who obtained abortions with both pregnant women who had a child and women who sought abortions but were found after being tested not to be pregnant. They found that the treated women (with abortions) had a better educational outcome after 2 years than either of the two control groups, which showed educational outcomes that were about equal to each other. If the only control group had been the women who had children, the better educational outcomes found in women who had abortions could be challenged as being due to subsequent child care demands on the control group mothers rather than to the abortion itself. But that argument could not apply to the control group of women who turned out not to be pregnant and so did not have increased child care demands. These results make it more difficult to argue that abortion will *decrease* subsequent educational attainment.

Improving the Posttest-Only Design Using a Predicted Interaction

Sometimes substantive theory is good enough to generate *a highly differentiated causal hypothesis* that, if corroborated, would rule out many internal validity threats because they are not capable of generating such complex empirical implications. An example is Seaver's (1973) quasi-experiment on the effects of teacher performance expectancies on students' academic achievement. Seaver located children whose older siblings had previously obtained high (or low) grades and achievement scores in school. He divided these two (high- versus low-achieving) groups into those who had the same teacher their older sibling had had and those who had a different teacher. Seaver predicted that teacher expectancies should cause children with high-performing siblings to outperform children with low-performing siblings by a greater amount if they had the same teachers rather than different ones. The data corroborated this predicted statistical interaction. In this case, predicted interactions were useful because (1) substantive theory predicted a complex data pattern, (2) sibling control groups were available that, although nonequivalent, were plausibly similar on many family background factors, (3) outcome measures (academic achievement) were reliably measured, and

(4) large sample sizes allowed a powerful test of the interaction. These circumstances are rare in social research, but they illustrate the role that coherence of findings can play in strengthening an inference.

However, the researcher should be cautious even when a predicted interaction is confirmed—Reichardt (1985) showed that results from the Seaver study may be due to a regression artifact. When Seaver partitioned students into four groups, he also partitioned teachers. Suppose the older siblings with above-average performance had teachers whose abilities (not expectations) were above average. A younger sibling assigned to the same teacher would then receive above-average teaching; but a younger sibling assigned to a different teacher might receive closer to average teaching. A similar argument applies for older siblings who performed poorly and whose teachers may have been less able. Differences in teacher effectiveness might thus account for the crossover interaction Seaver obtained. But other internal validity threats still seem implausible, and this threat could be assessed empirically if measures of teacher ability were available.

Improving Designs Without Control Groups by Constructing Contrasts Other Than With Independent Control Groups

When it is not possible to gather prospective data on the kinds of independent control groups just discussed, it is sometimes possible to construct contrasts[9] that try to mimic the function of an independent control group. Three such contrasts are (1) a regression extrapolation that compares actual and projected posttest scores, (2) a normed comparison that compares treatment recipients to normed samples, and (3) secondary data that compares treatment recipients to samples drawn from previously gathered data, such as population-based surveys. All of these possibilities have great weaknesses, so our preference is that these constructed contrasts not be used by themselves. However, they are often inexpensive and convenient, and so a study can afford to combine them with other quasi-experimental design features to shed light on remaining alternative explanations at little added cost.

Regression Extrapolation Contrasts

This design compares the obtained posttest score of the treatment group with the score it was predicted to obtain based on other information. For example, Cook et al. (1975) studied the effects of viewing the television program *Sesame Street* in several parts of the country. A pretest was administered, followed 6 months later by a posttest. Age (in months) at pretest was then used to predict pretest academic achievement, resulting in an estimate of how much achievement gain is expected for

[9] We call these contrasts because they would not generally be regarded as true control groups, even though the contrast of their results to those of the treatment group can sometimes help support a counterfactual inference.

each month a child aged. This monthly change estimate was then used to predict how much each child would have gained due to maturation in the 6 months separating pretest from posttest. The resulting prediction (which could not have been influenced by viewing *Sesame Street* because it was restricted to pretest measures) was then compared against the observed posttest data (which presumably was influenced by viewing *Sesame Street*). The regression equation generating the monthly gain estimate could also include measures of other threats to validity, such as parental socioeconomic status or other measures of selection bias.

However, by itself this approach has many severe problems. Without full knowledge of all threats to validity, predicted scores will rarely yield valid counterfactual inferences.[10] Moreover, the analysis depends on stable estimation using reliable measures and large samples. It also cannot address the possibility of history causing spurious effects after the pretest. Next, a testing artifact is possible because the obtained posttest is based on a second testing, whereas the first one obviously is not. And finally, this form of analysis is often used with aggregated school data to see if a school is doing better in a particular year than would be predicted for it on the basis of past academic performance and the nature of the student body and perhaps even the faculty. To do this well requires the use of multilevel data analytic approaches (Raudenbush & Willms, 1995; Willms, 1992) that respect the fact that individual responses are nested within schools. In our judgment, regression extrapolation contrasts are only worth doing when no other form of control group is possible or as an adjunct to a larger design. Indeed, Cook et al. (1975) used it as only one of many probes of the hypotheses about *Sesame Street*'s effectiveness, some of the others being much more conventional.

Normed Comparison Contrasts

In this case, obtained performance of the treatment group at pretest and posttest is compared with whatever published norms are available that might shed light on a counterfactual inference of interest. For example, Jacobson et al. (1984) compared couples treated for marital distress with norms for well-adjusted couples on the Marital Adjustment Scale to see if marital therapy made couples well-adjusted. Similarly, Nietzel, Russell, Hemmings, and Gretter (1987) reviewed studies that compared groups receiving therapy for depression with norms on the Beck Depression Inventory (Beck, Ward, Mendelsohn, Mock, & Erbaugh, 1961) to see if treated individuals eventually achieved levels of well-being similar to those reported by nondepressed adults. In each of these cases, the posttest score of the treated group was compared with the norm, with effects judged clinically significant if the treated group met or exceeded the normative standard.

This form of comparison is also routinely used in educational studies to assess whether a group of students, classrooms, or schools rises over time in its per-

[10] The issues raised here are the same as those raised in selection bias modeling and structural equation causal modeling covered in the appendix to Chapter 5, so we defer more detailed discussion to that chapter.

centile ranking on some published test. The possible rankings are taken from published norms and are meant to reflect the students' performance and its change relative to how students in the original norming sample did. Studies using this method are severely limited. The normative contrast is a very weak counterfactual, providing little indication of how the actual treated participants would have performed without treatment. In fact, to the extent that the normative group is selected for being superior to the treated group, it becomes impossible for this comparison to reflect the treatment effect even if treatment improves outcomes greatly compared with what would have occurred without treatment. A treatment that is highly effective compared with a standard control group would often be judged ineffective against such a norm. The normative comparison group is also threatened by selection, because the normed sample usually differs from the treated sample; by history, because the norms were usually gathered well before data from the treated sample were; by testing, if the treated sample was pretested but the normative sample was not; by regression, if the treated sample was chosen because of high need (or high merit) but the normative sample was not; by instrumentation, if the conditions of measurement were much different for treated and normative groups; and by maturation, if the treated group was rapidly changing compared with the normative group. These threats can sometimes be ameliorated by using local norms gathered on the same population as the treated group, by ensuring that the conditions and timing of testing are similar over groups, and by selecting normative samples thought to have similar maturational experiences.

Secondary Source Contrasts

Even without published norms, researchers can sometimes construct opportunistic contrasts from secondary sources. For instance, medical researchers sometimes use clinical series, records of cases treated prior to a new treatment, for this purpose; medical data registers of all patients with a particular condition are similarly used; and historical controls from within the same institution are sometimes available (D'Agostino & Kwan, 1995). Studies of the effects of substance abuse prevention programs have supplemented their basic one-group designs with data from national or state surveys (Furlong, Casas, Corral, & Gordon, 1997; Shaw, Rosati, Salzman, Coles, & McGeary, 1997). Labor economists use national data sets this way, using current population surveys or panel studies on income dynamics to create contrasts against which to evaluate job training programs. These contrasts are all useful as preliminary indications of the possibility of treatment effects—for example, in Phase II trials in medicine to establish if a new treatment has promise.

However, any such use of archival or historical data faces daunting practical obstacles, including those just described for normative samples. The data may have been collected for different reasons from those that motivate the present study, despite superficial similarities in published descriptions, which may reduce comparability. Data may be of insufficient quality; for example, reliability checks used in treatment data may not have been used in the archival data, and missing data may be prevalent. The data may not include covariates that are needed to adjust for group differences and to diagnose group comparability. Such contrasts can add significant bias to effect estimates (Sacks, Chalmers, & Smith, 1982, 1983) because of changes in the populations who have a problem (e.g., the populations with AIDS have changed over time) or who are eligible for or have access to treatment (e.g., changes through which AIDS treatments are reimbursable). Again, then, the use of these secondary source contrasts best serves causal inference goals when coupled with other design features to create a more complex quasi-experiment.

The Case-Control Design

In the designs considered so far, participants are divided into groups that do or do not receive a treatment, and their subsequent outcomes are examined. This search for the effects of causes is characteristic of experiments. However, sometimes it is not feasible or ethical to experiment. For instance, in the 1960s the drug diethylstilbestrol (DES) was given to pregnant women who were at risk of miscarriage because of bleeding. DES became suspected of causing vaginal cancer in their daughters. It would be unethical to knowingly give this drug to some women and not to others to see if it caused cancer. In addition, vaginal cancer is so rare and takes so long to develop that enormous sample sizes and many years would be needed to reliably detect an experimental outcome. In such cases, an option is to use the **case-control design** (also called **case-referent**, case-comparative, case-history, or retrospective design) that was invented and is widely used in epidemiology. In this design, one group consists of cases that have the outcome of interest, and the other group consists of controls that do not have it.[11] The outcome in this design is typically dichotomous, such as pass or fail, diseased or healthy, alive or dead, married or divorced, smoke-free or smoking, relapsed or drug-free, depressed or not depressed, or improved or not improved. Cases and controls are then compared using retrospective data to see if cases experienced the hypothesized cause more often than controls. Herbst, Ulfelder, and Poskanzer (1971) identified 8 cases with vaginal cancer and matched them to 32 controls without vaginal cancer who had been born within 5 days of the case at the same hospital. Seven of eight cases had received DES, but none of the controls had.

The case-control design is excellent for generating hypotheses about causal connections. Causal relationships first identified by case-control studies include smoking and cancer, birth control pills and thromboembolism, and DES and vaginal cancer (Vessey, 1979). Case-control studies are more feasible than experiments in cases in which an outcome is rare or takes years to develop; they are often

[11] We include this design in this chapter rather than the next one because case-control studies do not typically use a pretest, though it is possible to do so. The term "control" can be a bit misleading in case-control designs. In experiments, it implies a group that was not exposed to treatment. In case-control studies, however, some members of the control group may have received treatment even though they did not development the condition of interest.

cheaper and logistically easier to conduct; they may decrease risk to participants who could be needlessly exposed to a harmful experimental treatment; and they allow easy examination of multiple causes of a condition (Baker & Curbow, 1991).

Certain methodological problems are typical of case-control studies. The definition and selection of cases requires a decision about what counts as the presence or absence of the outcome of interest. But the relevant community of scholars may disagree about that decision; and even if they do agree, methods for assessing the outcome may be unreliable or of low validity. Moreover, definitions and measures may change over time, and so be different for cases diagnosed recently compared with cases diagnosed many years ago. Even when cases are diagnosed contemporaneously, they are often identified because their outcome brings them to the attention of a treatment source, as when women with vaginal cancer present for diagnosis and treatment. Controls rarely present this way, as they did not have the outcome. Hence the selection mechanism inevitably differs for the two groups. Attrition occurs when the outcome causes some cases to be unavailable (e.g., death from cancer occurs before the study begins). Those missing cases may have different characteristics than the available cases, a difference that can cause bias if the distribution of controls is not similarly truncated.

Selection of control cases is difficult. A common method is to choose controls to represent a general population. Randomly sampled controls are the exemplar. But when random sampling is not feasible, the control is often chosen by matching controls to cases on characteristics related to outcome. Neighborhood controls (exposure to a similar living environment) and hospital controls (exposure to the same facility) are common kinds of matched controls (Lund, 1989). Herbst et al.'s (1971) method of selecting controls from those born in the same hospital at the same time as the cases presumably increased the similarity of cases and controls on geographic influences, demographics, and birth cohort effects. However, matched controls still differ from cases in unobserved ways that can be confounded with the presumed cause and can be the actual cause of the outcome. For example, one study of children with diabetes used "friendly controls" in which parents of cases provided names of two age- and sex-matched friends of the children as controls (Siemiatycki, Colle, Campbell, Dewar, & Belmonte, 1989). Results suggested that children with diabetes were more likely to have school problems, few friends, trouble sleeping, hospitalizations and accidents, recent bereavements, and parental divorce. But are these causes or confounds? It turned out that parents tended to nominate sociable people as their children's friends, so the controls were disproportionately positive on social variables (Siemiatycki, 1989). The use of multiple controls can help avoid such problems (Kleinbaum, Kupper, & Morgenstern, 1982; Rosenbaum, 1995a)—a group from the same source of care as the cases, another from the same neighborhood, and a third sampled randomly from the general population (Baker & Curbow, 1991; Lund, 1989). Differences over controls in estimates of the causal relationship help to index the amount of hidden bias that may be present.

Further, which control population is most relevant depends greatly on the desired inference. For example, the use of a case's neighbors as controls is common. But if the question was whether traveler's diarrhea observed in U.S. citizens in a Mexican hospital is caused by drinking tequila, using neighbors would be inappropriate compared with using other non-Mexican travelers in the same hospital's catchment area (Miettinen, 1985). A general population control might be appropriate if little is known about specific causes, but a more narrowly defined control might be useful if the causal question is highly specific (Garber & Hollon, 1991). Also, cases that share the same problem are not always homogenous in the cause of that problem. For example, if cases are patients with a staphylococcus infection in a hospital, some may have contracted the infection outside the hospital and others within it (iatrogenic infection), requiring different controls.

Assessment of treatment exposure in case-control designs is retrospectively reconstructed from fallible sources such as memory or records. Hence classification as either having had exposure to treatment or not having had exposure is rarely fully accurate. Cases may have more incentive than controls to remember exposure to risk factors; for example, they may perceive that their diagnostic accuracy depends on an accurate history. Further, exposure to treatment is almost certainly confounded with other covariates. In the DES example, mothers were given DES because of increased risk of miscarriage due to bleeding, so the latter risk is confounded with exposure to DES. In this case that risk is implausible as a cause of vaginal cancer, given other correlational research supporting the link between DES and vaginal cancer, given randomized animal trials showing the effect, and given the lack of a theoretical expectation to link miscarriage risk and vaginal cancer (Potvin & Campbell, 1996). Still, exposure to treatment is usually confounded in case control studies, making the causal connection between treatment and outcome more tenuous.

These examples all illustrate that, because a case-control study probes causal inferences, the logic of ruling out threats to validity applies to it (Potvin & Campbell, 1996; Campbell & Russo, 1999). In fact, a literature about threats to validity with the case-control design has grown independently of the tradition presented in this book. Sackett (1979) lists these threats (Table 4.3)—although not all the items listed are necessarily sources of bias (e.g., the exclusion of outliers now has more extensive justification than it did 20 years ago). A large literature has developed to improve causal inferences from case-control studies, especially regarding analyses to adjust for potential confounds (Ahlbom & Norell, 1990; Greenland & Robins, 1986; Kleinbaum et al., 1982; Rothman, 1986; Schlesselman, 1982). Still, we believe that the case-control design deserves more widespread use in areas other than public health, although more for its ability to generate causal hypotheses than to test them well.

TABLE 4.3 Threats to Validity in Case-Control Studies

1. In *reading up* on the field:
    - a. *The biases of rhetoric.* Any of several techniques used to convince the reader without appealing to reason.
    - b. *The all's well literature bias.* Scientific or professional societies may publish reports or editorials that omit or play down controversies or disparate results.
    - c. *One-sided reference bias.* Authors may restrict their references to only those works that support their position: a literature review with a single starting point risks confinement to a single side of the issue.
    - d. *Positive results bias.* Authors are more likely to submit, and editors accept, positive than null results.
    - e. *Hot stuff bias.* When a topic is hot, neither investigators nor editors may be able to resist the temptation to publish additional results, no matter how preliminary or shaky.
2. In *specifying and selecting* the study sample:
    - a. *Popularity bias.* The admission of patients to some practices, institutions, or procedures (surgery, autopsy) is influenced by the interest stirred up by the presenting conditions and its possible causes.
    - b. *Centripetal bias.* The reputations of certain clinicians and institutions cause individuals with specific disorders or exposures to gravitate toward them.
    - c. *Referral filter bias.* As a group of ill are referred from primary to secondary to tertiary care, the concentration of rare causes, multiple diagnoses, and "hopeless cases" may increase.
    - d. *Diagnostic access bias.* Individuals differ in their geographic, temporal, and economic access to the diagnostic procedures that label them as having a given disease.
    - e. *Diagnostic suspicion bias.* A knowledge of the subject's prior exposure to a putative cause (ethnicity, taking a certain drug, having a second disorder, being exposed in an epidemic) may influence both the intensity and the outcome of the diagnostic process.
    - f. *Unmasking (detection signal) bias.* An innocent exposure may become suspect if, rather than causing a disease, it causes a sign or symptom that precipitates a search for the disease.
    - g. *Mimicry bias.* An innocent exposure may become suspect if, rather than causing a disease, it causes a (benign) disorder that resembles the disease.
    - h. *Previous opinion bias.* The tactics and results of a previous diagnostic process on a patient, if known, may affect the tactics and results of a subsequent diagnostic process on the same patient.
    - i. *Wrong sample size bias.* Samples that are too small can prove nothing; samples that are too large can prove anything.
    - j. *Admission rate (Berkson) bias.* If hospitalization rates differ for different exposure/disease groups, the relation between exposure and disease will become distorted in hospital-based studies.
    - k. *Prevalence-incidence (Neyman) bias.* A late look at those exposed (or affected) early will miss fatal and other short episodes, plus mild or "silent" cases and cases in which evidence of exposure disappears with disease onset.
    - l. *Diagnostic vogue bias.* The same illness may receive different diagnostic labels at different points in space or time.
    - m. *Diagnostic purity bias.* When "pure" diagnostic groups exclude co-morbidity, they may become non-representative.

TABLE 4.3 Continued

n. *Procedure selection bias.* Certain clinical procedures may be preferentially offered to those who are poor risks.

o. *Missing clinical data bias.* Missing clinical data may be missing because they are normal, negative, never measured, or measured but never recorded.

p. *Non-contemporaneous control bias.* Secular changes in definitions, exposures, diagnosis, diseases, and treatments may render non-contemporaneous controls non-comparable.

q. *Starting time bias.* The failure to identify a common starting time for exposure or illness may lead to systematic misclassifications.

r. *Unacceptable disease bias.* When disorders are socially unacceptable (V.D., suicide, insanity), they tend to be under-reported.

s. *Migrator bias.* Migrants may differ systematically from those who stay home.

t. *Membership bias.* Membership in a group (the employed, joggers, etc.) may imply a degree of health that differs systematically from that of the general population.

u. *Non-respondent bias.* Non-respondents (or "late comers") from a specified sample may exhibit exposures or outcomes that differ from those of respondents (or "early comers").

v. *Volunteer bias.* Volunteers or "early comers" from a specified sample may exhibit exposures or outcomes (they tend to be healthier) that differ from those of non-volunteers or "late comers."

3. In *executing* the experimental manoeuvre (or exposure):

a. *Contamination bias.* In an experiment when members of the control group inadvertently receive the experimental manoeuvre, the difference in outcomes between experimental and control patients may be systematically reduced.

b. *Withdrawal bias.* Patients who are withdrawn from an experiment may differ systematically from those who remain.

c. *Compliance bias.* In experiments requiring patient adherence to therapy, issues of efficacy become confounded with those of compliance.

d. *Therapeutic personality bias.* When treatment is not "blind," the therapist's convictions about efficacy may systematically influence both outcomes (positive personality) and their measurement (desire for positive results).

e. *Bogus control bias.* When patients who are allocated to an experimental manoeuvre die or sicken before or during its administration and are omitted or re-allocated to the control group, the experimental manoeuvre will appear spuriously superior.

4. In *measuring* exposures and outcomes:

a. *Insensitive measure bias.* When outcome measures are incapable of detecting clinically significant changes or differences, Type II errors occur.

b. *Underlying cause bias (rumination bias).* Cases may ruminate about possible causes for their illnesses and thus exhibit different recall or prior exposures than controls.

c. *End-digit preference bias.* In converting analog to digital data, observers may record some terminal digits with an unusual frequency.

d. *Apprehension bias.* Certain measures (pulse, blood pressure) may alter systematically from their usual levels if the subject is apprehensive.

e. *Unacceptability bias.* Measurements that hurt, embarrass, or invade privacy may be systematically refused or evaded.

f. *Obsequiousness bias.* Subjects may systematically alter questionnaire responses in the direction they perceive desired by the investigator.

TABLE 4.3 Continued

g. *Expectation bias.* Observers may systematically err in measuring and recording observation so that they concur with prior expectations.

h. *Substitution game.* The substitution of a risk factor that has not been established as causal for its associated outcome.

i. *Family information bias.* The flow of family information about exposure and illness is stimulated by, and directed to, a new case in its midst.

j. *Exposure suspicion bias.* A knowledge of the subject's disease status may influence both the intensity and outcome of a search for exposure to the putative cause.

k. *Recall bias.* Questions about specific exposures may be asked several times of cases but only once of controls. (See also the *underlying cause bias.*)

l. *Attention bias.* Study subjects may systematically alter their behavior when they know they are being observed.

m. *Instrument bias.* Defects in the calibration or maintenance of measurement instruments may lead to systematic deviation from true values.

5. In *analyzing* the data:

a. *Post-hoc significance bias.* When decision levels or "tails" for *x* and *B* are selected *after* the data have been examined, conclusions may be biased.

b. *Data dredging bias (looking for the pony).* When data are reviewed for all possible associations without prior hypothesis, the results are suitable for hypothesis-forming activities only.

c. *Scale degradation bias.* The degradation and collapsing of measurement scales tends to obscure differences between groups under comparison.

d. *Tidying-up bias.* The exclusion of outliers or other untidy results cannot be justified on statistical grounds and may lead to bias.

e. *Repeated peeks bias.* Repeated peeks at accumulating data in a randomized trial are not dependent and may lead to inappropriate termination.

6. In *interpreting* the analysis:

a. *Mistaken identity bias.* In compliance trials, strategies directed toward improving the patient's compliance may, instead or in addition, cause the treating clinician to prescribe more vigorously; the effect upon achievement of the treatment goal may be misinterpreted.

b. *Cognitive dissonance bias.* The belief in a given mechanism may increase rather than decrease in the face of contradictory evidence.

c. *Magnitude bias.* In interpreting a finding, the selection of a scale of measurement may markedly affect the interpretation.

d. *Significance bias.* The confusion of statistical significance, on the one hand, with biology or clinical or health care significance, on the other hand, can lead to fruitless studies and useless conclusions.

e. *Correlation bias.* Equating correlation with causation leads to errors of both kinds.

f. *Under-exhaustion bias.* The failure to exhaust the hypothesis space may lead to authoritarian rather than authoritative interpretation.

Source: Reprinted from "Bias in analytic research," by D. L. Sackett, 1979, *Journal of Chronic Diseases, 32,* pp. 51–63. Copyright 1979 by Elsevier Science. Reprinted with permission.

CONCLUSION

This chapter discusses and criticizes a wide variety of quasi-experimental designs that are often used but that frequently provide a weak basis for causal inference compared with the designs that follow in this book. These designs are weak primarily because they lack either a pretest or a control group. Nonetheless, these designs have a place in the methodologist's repertoire when many alternative causes are implausible on practical or theoretical grounds, when uncertainty reduction about cause is a low priority, and when the need is to generate causal hypotheses for further study with stronger designs. Fortunately, these designs can be strengthened by adding more design elements that are selected to address concerns about particularly salient threats to internal validity, particularly by adding both pretests and control groups, a design to which we now turn.