CHAPTER THREE

# Causal Inference and Generalization in Field Settings

## Experimental and Quasi-Experimental Designs

STEPHEN G. WEST, JEREMY C. BIESANZ, AND STEVEN C. PITTS

The purpose of this chapter is to introduce researchers in social psychology to designs that permit relatively strong causal inferences in the field. We begin the chapter by considering some basic issues in inferring causality, drawing on work by Rubin and his associates (Holland, 1986, 1988; Rubin, 1974, 1978) in statistics and by Campbell and his associates (Campbell, 1957; Campbell & Stanley, 1966; Cook, 1993; Cook & Campbell, 1979; Shadish, Cook, & Campbell, in press) in psychology. Rubin's approach emphasized formal statistical criteria for inference; Campbell's approach emphasized concepts from philosophy of science and the practical issues confronting social researchers. These approaches are then applied to provide insights on a variety of difficult issues that arise in randomized experiments when they are conducted in the field. In addition, Cook's (1993) new perspective on the generalization of causal effects is also discussed. We then turn to consideration of three classes of quasi-experimental designs – the regression discontinuity design, the interrupted time series design, and the nonequivalent control group design – that can sometimes provide a relatively strong basis for causal inference. The application of the Rubin and the Campbell frameworks helps identify strengths and weaknesses of each design. Methods of strengthening each design type with respect to causal inference and generalization of causal effects are also considered.

Stephen G. West was partially supported by NIMH Grant P50-MN39246 during the writing of this chapter. We thank Tom Belin, Tom Cook, Bill Graziano, Chick Judd, Chip Reichardt, Harry Reis, Will Shadish, and the graduate students of the Fall 1997 Psychology 555 class (experimental and quasi-experimental designs in research) at Arizona State University for their comments and suggestions on an earlier version of this chapter. Correspondence should be directed to Stephen G. West, Department of Psychology, Arizona State University, Tempe, AZ 85287-1104 (e-mail: sgwest@asu.edu).

The emphasis on designs for field research in the present chapter contrasts sharply with recent practice in basic social psychology in which the modal research design consists of a randomized experiment, conducted in the laboratory, lasting no more than 1 hour, and using undergraduate students as the participants (West, Newsom, & Fenaughty, 1992). This practice has clear strengths that are articulated in other chapters in this handbook. At the same time, the recent extraordinary level of dominance of this practice potentially limits the generalization of social psychological findings in important ways. At earlier points in the history of the field, it was far easier to find examples of basic social psychological hypotheses that were tested in both laboratory and field settings (see Bickman & Henchy, 1972; Swingle, 1973, for collections of examples); currently, the majority of research conducted by social psychologists in field settings is applied in nature. The earlier interplay between laboratory and field research often provided a more convincing basis for generalization of causal effects found in basic social psychological investigations. We suggest that perhaps at least some of the current widespread perception that many articles in our leading journals "just aren't that interesting anymore" (Reis & Stiller, 1992, p. 470; see also Funder, 1992) may be occasioned by the limited contexts in which social psychological phenomena are studied and the distance of present research paradigms from the real world.

The current modal laboratory research paradigm is particularly good at establishing internal validity, that is, the treatment caused the observed response. However, as complications are introduced into the

laboratory setting, establishing internal validity becomes more tenuous. Experiments conducted over repeated sessions involve attrition; requests for participants to bring peers or parents to the laboratory are not always fulfilled. Basic and applied social psychological research conducted in field settings may further magnify these potential issues. Consequently, researchers must increase their focus on articulating those threats to internal validity, which Campbell (1957) termed "plausible rival hypotheses," that may be problematic in their specific research context. A central theme of this chapter is the identification of major classes of plausible rival hypotheses associated with each of the design types and the use of specific design and analysis strategies that can help rule them out. And, as we will indicate, many of these strategies also turn out to have potential implications for traditional laboratory experiments as well.

## A FRAMEWORK FOR CAUSAL INFERENCE IN EXPERIMENTS AND QUASI-EXPERIMENTS

Over the past two decades, Donald Rubin and his colleagues in statistics (Angrist, Imbens, & Rubin, 1996; Holland, 1986, 1988; Imbens & Rubin, 1997; Rosenbaum & Rubin, 1983, 1984; Rubin, 1974, 1978, 1986) have developed a very useful framework for understanding the causal effects of treatments. This framework has come to be known as the Rubin Causal Model (RCM). The use of the RCM is particularly helpful in identifying strengths and limitations of many of the experimental and quasi-experimental designs in which the independent variable is manipulated and posttest measures are collected at only one point in time following the treatment.

Consider the simple case of two treatments whose effects the researcher wishes to compare. For example, the researcher may wish to compare a severe frustration (treatment condition) and a mild frustration (comparison condition) in the level of aggressive responses they produce (see Berkowitz, 1993). Or, the researcher may wish to compare a 10-week smoking prevention program (treatment condition) and a no program group (comparison condition) on attitudes toward smoking among teenagers (see Flay, 1986).

Rubin begins with a consideration of the hypothetical ideal conditions under which a causal effect could be observed. He defines the causal effect as the difference between what would have happened to the participant under the treatment condition and what *would* have happened to the same participant under the control condition under identical circumstances. That is, the causal effect is defined as:

$$Y_t(u) - Y_c(u).$$

Here, $t$ refers to the treatment condition, $c$ refers to the comparison condition (often a control group), $Y$ is the observed response (dependent measure), and $u$ is the unit (in psychology, typically a specific participant) on which we observe the effects of the two treatments.

Rubin's definition leads to a clear theoretical statement of what a causal effect is, but it also implies a fundamental problem. "It is impossible to observe the value of $Y_t(u)$ and $Y_c(u)$ on the same unit and, therefore, it is impossible to *observe* the effect of $t$ on $u$" (Holland, 1986, p. 947, italics in original). We cannot expose a preteenage boy to the 10-week smoking prevention program, measure his attitudes toward smoking, then return the child to the beginning of the school year, expose him to the comparison (control) program and remeasure his attitudes toward smoking. Because of this fundamental problem of causal inference, we will not be able to observe causality directly. However, by making specific assumptions, we can develop research designs that permit us to infer causality. The certainty of the causal inference will depend strongly on the viability of the assumptions that we make. Three design approaches that address the fundamental problem of causal inference are available, which may be employed alone or in combination. After briefly presenting the first two design approaches, we focus on the third approach, randomization, because it is most often used in basic and applied social psychology.

### Design Approaches to the Fundamental Problem of Causal Inference

**WITHIN-SUBJECT DESIGNS.** In within-subject designs, participants are exposed to treatment (i) and their responses are measured, following which they are exposed to treatment (ii) and their responses are measured. Within-subject designs make two strong assumptions. First is *temporal stability*, which means that the response does not depend on the time the treatment is delivered. Many factors such as normal human development, historical events, fatigue, or even daily or monthly cycles (see Cook & Campbell, 1979, chapter 2) can affect participants' responses. Second is *causal transience*, which means that the effects of each treatment and each measurement of the response do not persist over time. For example, in a study of the effects of a high-stress and a low-stress film clip on increases in blood pressure, the researcher would need to make the strong assumptions that exposure to the initial film

clip did not affect the perception of the second clip, the participant had fully returned to her baseline level for all bodily systems (i.e., not only blood pressure, but also physiological, motivational, and cognitive systems), and that the initial blood pressure measurement did not affect subsequent blood pressure measurements (e.g., changes in blood pressure readings from the participant's anticipation of the unpleasant constriction from the inflation of the blood pressure cuff). For studies of physical devices, such as the performance of engines under varying treatment conditions, the assumptions of temporal stability and causal transience will often be reasonable. For studies of social psychological phenomena, these assumptions will often be more problematic. Greenwald (1976), Erlebacher (1977), Rosenthal and Rubin (1980), Judd and Kenny (1981), and Keren (1993) presented fuller discussions of issues in within-subject designs.[1]

**UNIT HOMOGENEITY.** If experimental units can be assumed to be identical in all respects, then it makes no difference which unit receives the treatment. That is, $Y_t(u_1) = Y_t(u_2)$ and $Y_c(u_1) = Y_c(u_2)$, where $u_1$ and $u_2$ represent two different units (participants). In this case, $Y_t(u_1) - Y_c(u_2)$ [or $Y_t(u_2) - Y_c(u_1)$] will provide an accurate estimate of the causal effect. Again, this assumption of unit homogeneity is often made in the physical sciences and engineering, but it will almost never be reasonable in social psychology. Even monozygotic (identical) twins who are known to share identical genes can easily differ in knowledge, motivation, mood, or other participant-related factors that may affect their responses to treatment.

**RANDOMIZATION.** Randomization is used to equate approximately the treatment and control groups at pretest, prior to any treatment delivery. In randomization, participants are assigned to treatment conditions using a method that gives every participant an equal chance of being assigned to the treatment and control conditions.[2] Common randomization methods include flipping a coin (e.g., heads = treatment; tails = control), drawing a number out of a hat (e.g., 1 = treatment; 2 = control), or using a computer to generate random numbers identifying treatment and control group assignments. Although randomization is normally straightforward in the laboratory, some field research settings present very difficult randomization problems. For example, in health care settings participants often arrive when they become ill; facilities to deliver the experimental treatment, the control treatment, or both may or may not be available at that time. A wide variety of different methods reviewed by Boruch (1997) and Shadish et al. (in press) have been developed to address complex randomization problems.

Following random assignment to treatment conditions, each participant then receives the treatment condition (e.g., experimental treatment vs. comparison [control] treatment) to which he or she was assigned. Then, the responses of each participant are measured after the receipt of the treatment condition. These procedures are characteristic of the basic randomized experiment used in both laboratory and field settings in social psychology.

Random assignment means that the variable representing the treatment condition (treatment vs. control) can be expected on average to be independent of any measured or unmeasured variable prior to treatment. This outcome is reflected in two closely related ways.

1. At pretest, prior to any treatment delivery, the mean levels in the treatment and control groups will, on average, be equal for any measured or unmeasured variable. More formally, the expected values for the two groups (Group 1 and Group 2) will be equal, $E(\bar{Y}_1) = E(\bar{Y}_2)$, for any variable $Y$. At pretest, the demographic characteristics, the attitudes, the motivations, the personality traits, the abilities, as well as any other participant attributes can, on average,

[1] A hybrid approach that combines elements of the within-subject design and randomization approaches has been traditionally used in some areas of experimental psychology. In the simplest version of such a design, half of the participants are randomly assigned to receive treatment (i) and then treatment (ii), whereas the other half of the participants are randomly assigned to receive treatment (ii) and then treatment (i). Such designs make the typically untestable assumption that there are no interactions between treatment condition and order of presentation or treatment condition and trial. This assumption will be met when there is no causal transience. Statistical presentations of the assumptions and analysis of such counterbalanced and Latin Square designs can be found in Winer (1971) and Kirk (1995).

[2] In fact, the probability of being assigned to the treatment and control conditions may differ. For example, the probability of assignment to the treatment group might be .25 and the probability of assignment to the control group might be .75 for each participant. Unequal allocation of participants to treatment and control groups is normally used when the cost or difficulty in implementing one of the treatment conditions is substantially greater than for the other. We will assume equal allocation of participants to treatment and control groups here. Given usual statistical assumptions underlying hypothesis testing, this equal allocation strategy maximizes the power of the test of the null hypothesis.

be expected to be equal in the treatment and control groups.

2. The treatment variable (hereafter referred to as $t$ for treatment and $c$ for comparison) will, on average, be unrelated to any measured or unmeasured participant variable prior to treatment. One implication of this is that the correlation of the treatment variable and any participant variable measured prior to treatment delivery will, on average, be 0.

These two outcomes allow the RCM to provide a new definition of the causal effect that is appropriate for randomized experiments. The estimate of the causal effect is now defined as:

$$\bar{Y}_t - \bar{Y}_c.$$

Three observations are in order. First, the comparison must now shift to the *average* response for the group of participants receiving the experimental treatment compared with the *average* response for the group of participants receiving the control treatment. Causal inferences may no longer be made about individual participants. Second, the two outcomes are expectations: They are what occurs "on average." Exact equivalence of the pretest means of variables and exact independence of the treatment and pretest measures of participant variables does *not* occur routinely in practice. With simple randomization, exact pretest equivalence and exact independence of the treatment and background variables occur only if randomization is applied to a very large ($n \rightarrow \infty$) population or the results are averaged across a very large number (all possible) of different randomizations of the same sample. In any given real sample, there is no guarantee that pretest means will *not* differ on important variables, in which case the estimate of the causal effect, $\bar{Y}_t - \bar{Y}_c$, will be too low or too high. Randomization replaces definitive statements about causal effects with probabilistic statements based on sampling theory. Third, for causal inference we need to assume that neither the randomization process itself nor participants' potential awareness of other participants' treatment conditions influence the participants' responses. This assumption, known as the stable-unit-treatment-value assumption is considered in more detail following the presentation of an illustrative example to help make some of these ideas concrete.

## Illustrative Example: Randomization and Rubin's Causal Model

In Table 3.1 we present an example data set that we will use throughout this section of the chapter. There

**TABLE 3.1. Illustration: Estimating the Causal Effect – Ideal Case**

| Participant | $a_1$ | $a_2$ | $a_3$ | $Y_c$ | $Y_t$ |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | 1.5 |
| 2 | 1 | 0 | 0 | 2 | 2.5 |
| 3 | 1 | 0 | 1 | 2 | 2.5 |
| 4 | 1 | 0 | 0 | 2 | 2.5 |
| 5 | 0 | 0 | 0 | 2 | 2.5 |
| 6 | 0 | 1 | 0 | 3 | 3.5 |
| 7 | 1 | 1 | 0 | 3 | 3.5 |
| 8 | 1 | 0 | 1 | 3 | 3.5 |
| 9 | 0 | 1 | 1 | 3 | 3.5 |
| 10 | 0 | 1 | 0 | 3 | 3.5 |
| 11 | 0 | 1 | 1 | 3 | 3.5 |
| 12 | 1 | 0 | 0 | 4 | 4.5 |
| 13 | 0 | 0 | 1 | 4 | 4.5 |
| 14 | 0 | 1 | 1 | 4 | 4.5 |
| 15 | 1 | 1 | 1 | 4 | 4.5 |
| 16 | 0 | 0 | 1 | 5 | 5.5 |
| 17 | 1 | 1 | 1 | 1 | 1.5 |
| 18 | 1 | 1 | 0 | 2 | 2.5 |
| 19 | 0 | 0 | 0 | 2 | 2.5 |
| 20 | 1 | 0 | 1 | 2 | 2.5 |
| 21 | 1 | 1 | 0 | 2 | 2.5 |
| 22 | 0 | 1 | 0 | 3 | 3.5 |
| 23 | 1 | 1 | 1 | 3 | 3.5 |
| 24 | 1 | 1 | 1 | 3 | 3.5 |
| 25 | 0 | 0 | 1 | 3 | 3.5 |
| 26 | 0 | 0 | 0 | 3 | 3.5 |
| 27 | 1 | 1 | 1 | 3 | 3.5 |
| 28 | 1 | 0 | 0 | 4 | 4.5 |
| 29 | 1 | 0 | 0 | 4 | 4.5 |
| 30 | 0 | 0 | 1 | 4 | 4.5 |
| 31 | 0 | 0 | 1 | 4 | 4.5 |
| 32 | 0 | 0 | 0 | 5 | 5.5 |

*Note:* The column labeled $Y_c$ contains the true response of each participant in the control condition. The column labeled $Y_t$ contains the true response of each participant in the treatment condition. $a_1$, $a_2$, and $a_3$ represent three different random assignments of 16 participants to the control group and 16 participants to the treatment group. In each random assignment, 0 means the participant was assigned to the control group and 1 means the participant was assigned to the treatment group. The mean of all 32 participants under the control condition is 3.0, the mean of all 32 participants under the treatment condition is 3.5, and the standard deviation of each condition is 1.0. The distribution within each group is approximately normal. The causal effect for each participant is 0.5, corresponding to a moderate effect size.

are 32 participants. The example is constructed based on Rubin's ideal case: The response of each participant is observed under both the control ($Y_c$ column) and treatment ($Y_t$ column) conditions. The causal effect, $Y_t - Y_c$, has a constant size of 0.5 for each participant. The mean in the control group is 3.0, the mean in the treatment group is 3.5, and the standard deviation for the example is 1.0. The distributions within each group are roughly normal. Note that in this example the standardized measure of effect size is $d = \frac{\mu_t - \mu_c}{\sigma} = 0.50$. This standardized effect size is described by Cohen (1988) as moderate, and he believes it represents a typical effect size found in the behavioral sciences. By way of comparison, some recent meta-analyses have found mean effect sizes of $d = 0.54$ for the effect of alcohol on aggression (Ito, Miller, & Pollock, 1996), $d = 0.34$ for the effect of prevention programs on improved mental health outcomes for children (Durlak & Wells, 1997), and $d = .59$ for the relation between confidence and accuracy in studies of eyewitness identification (Sporer, Penrod, Read, & Cutler, 1995).

Also presented in Table 3.1 are three columns labeled $a_1$, $a_2$, and $a_3$. These represent three different actual random assignments of the 32 participants. In each case, 16 participants are assigned to the control condition, and 16 participants are assigned to the treatment condition. In general, there are $\binom{2n}{n}$ different possible random assignments, where $n$ is the number of participants in each treatment group (Cochran & Cox, 1957). In the present example, there are $\binom{32}{16} = \frac{32!}{16!16!} = \frac{32\times31\times30\times\cdots\times2\times1}{(16\times15\times14\times\cdots\times2\times1)(16\times15\times14\times\cdots\times2\times1)}$ combinations or over 600 million different possible random assignments.

The result of random assignment is that we will be able to observe only one of the two responses for each participant. Table 3.2 illustrates this result for randomization $a_1$. We see that for Participant 1, the response is observed under the control condition, but not the treatment condition; for Participant 2, the response is observed under the treatment, but not the control condition; and so on. The black squares represent the unobserved response for each participant.

In Table 3.3, we show the calculations of the causal effect for each of the three randomizations, $a_1$, $a_2$, and $a_3$. The example is constructed so that the true standard deviation ($\sigma = 1$) is known for both the treatment and control conditions for the full set of 32 participants in this ideal case, so we can use the $z$-test, rather than the more typical $t$-test, for two independent groups. The true causal effect of treatment in the population is 0.50. However, note that the estimates of the causal effect in the three samples are 0.0, −0.125, and 0.875. We can construct 95% confidence intervals (CI) around these estimates by applying the formula:

$$\text{CI} = (\bar{Y}_t - \bar{Y}_c) \pm (1.96)(SE),$$

where $SE$ is the standard error, which is equal to $\sqrt{\sigma^2(\frac{1}{n_t} + \frac{1}{n_c})}$. For randomization 1, CI = 0.0 ± (1.96)

**TABLE 3.2. Illustration: Estimating the Causal Effect – Randomized Experiment ($a_1$)**

| Participant | $a_1$ | $Y_c$ | $Y_t$ |
|---|---|---|---|
| 1 | 0 | 1 | ■ |
| 2 | 1 | ■ | 2.5 |
| 3 | 1 | ■ | 2.5 |
| 4 | 1 | ■ | 2.5 |
| 5 | 0 | 2 | ■ |
| 6 | 0 | 3 | ■ |
| 7 | 1 | ■ | 3.5 |
| 8 | 1 | ■ | 3.5 |
| 9 | 0 | 3 | ■ |
| 10 | 0 | 3 | ■ |
| 11 | 0 | 3 | ■ |
| 12 | 1 | ■ | 4.5 |
| 13 | 0 | 4 | ■ |
| 14 | 0 | 4 | ■ |
| 15 | 1 | ■ | 4.5 |
| 16 | 0 | 5 | ■ |
| 17 | 1 | ■ | 1.5 |
| 18 | 1 | ■ | 2.5 |
| 19 | 0 | 2 | ■ |
| 20 | 1 | ■ | 2.5 |
| 21 | 1 | ■ | 2.5 |
| 22 | 0 | 3 | ■ |
| 23 | 1 | ■ | 3.5 |
| 24 | 1 | ■ | 3.5 |
| 25 | 0 | 3 | ■ |
| 26 | 0 | 3 | ■ |
| 27 | 1 | ■ | 3.5 |
| 28 | 1 | ■ | 4.5 |
| 29 | 1 | ■ | 4.5 |
| 30 | 0 | 4 | ■ |
| 31 | 0 | 4 | ■ |
| 32 | 0 | 5 | ■ |

*Note.* The column labeled $Y_c$ contains the true response of each participant in the control condition. The column labeled $Y_t$ contains the true response of each participant in the treatment condition. $a_1$ represents the random assignment of 16 participants to the control group and 16 participants to the treatment group. ■ means response was not observed.

TABLE 3.3.

**Random assignment: $a_1$**

$$\bar{Y}_c = \frac{\sum Y_c}{n_c} = \frac{1+2+3+3+\cdots+4+4+5}{16} = \frac{52}{16} = 3.25$$

$$\bar{Y}_t = \frac{\sum Y_t}{n_t} = \frac{2.5+2.5+2.5+3.5+\cdots+3.5+4.5+4.5}{16} = \frac{52}{16} = 3.25$$

Estimate of causal effect: $\bar{Y}_t - \bar{Y}_c = 0.0$

$$z = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{\sigma^2\left(\frac{1}{n_t}+\frac{1}{n_c}\right)}} = \frac{0}{\sqrt{1\left(\frac{1}{16}+\frac{1}{16}\right)}} = 0$$

**Random assignment: $a_2$**

$$\bar{Y}_c = \frac{\sum Y_c}{n_c} = \frac{2+2+2+3+\cdots+4+4+5}{16} = \frac{53}{16} = 3.3125$$

$$\bar{Y}_t = \frac{\sum Y_t}{n_t} = \frac{1.5+2.5+3.5+3.5+\cdots+3.5+3.5+3.5}{16} = \frac{51}{16} = 3.1875$$

Estimate of causal effect: $\bar{Y}_t - \bar{Y}_c = -0.125$

$$z = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{\sigma^2\left(\frac{1}{n_t}+\frac{1}{n_c}\right)}} = \frac{-0.125}{\sqrt{1\left(\frac{1}{16}+\frac{1}{16}\right)}} = -0.35$$

**Random assignment: $a_3$**

$$\bar{Y}_c = \frac{\sum Y_c}{n_c} = \frac{1+2+2+2+\cdots+4+4+5}{16} = \frac{45}{16} = 2.8125$$

$$\bar{Y}_t = \frac{\sum Y_t}{n_t} = \frac{2.5+3.5+3.5+3.5+\cdots+3.5+4.5+4.5}{16} = \frac{59}{16} = 3.6875$$

Estimate of causal effect: $\bar{Y}_t - \bar{Y}_c = 0.875$

$$z = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{\sigma^2\left(\frac{1}{n_t}+\frac{1}{n_c}\right)}} = \frac{0.875}{\sqrt{1\left(\frac{1}{16}+\frac{1}{16}\right)}} = 2.47$$

(.354) = −.69 to +.69; for randomization 2, CI = −.82 to +.57; for randomization 3, CI = .18 to 1.57. None of these estimates represent really extreme values: Each of the 95% confidence intervals include the true causal effect of 0.50.

Also included in Table 3.3 are the traditional statistical tests of the significance for the null hypothesis that the true causal effect is 0. Two of these three significance tests fail to reject the null hypothesis at the traditionally accepted $p < .05$ value. Figure 3.1 presents the distribution of estimates of the causal effect that we would get if each of the over 600 million randomizations were performed, and we computed the causal effect $\bar{Y}_t - \bar{Y}_c$ for each randomization. Half of the estimates are less than 0.5 and about 8% are less than 0, surprisingly suggesting that the treatment might have

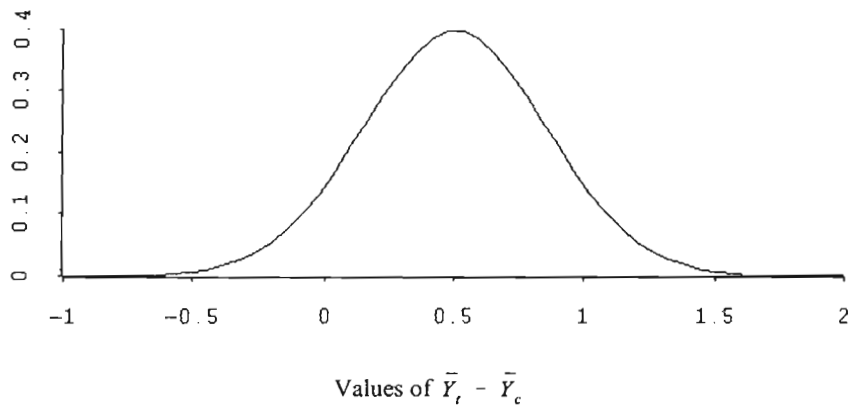



**Figure 3.1.** Distribution of $\bar{Y}_t - \bar{Y}_c$.

*Note.* The mean of the sampling distribution is 0.5. The standard deviation of the sampling distribution is 0.354. $\bar{Y}_t - \bar{Y}_c$ must exceed $1.96 * 0.354 = 0.69$ to reject the null hypothesis of no difference between $\mu_t - \mu_c$ in the population. The experimenter will correctly reject the null hypothesis about 30% of the time.

had a negative effect on the response.[3] Only about 30% of the estimates (i.e., those > 0.69) would lead to correct rejection of the null hypothesis of no causal effect. Indeed, had we been in the more usual situation and used a $t$-test rather than $z$-test because we did not know the population standard deviation, this value would have been slightly lower, about 28%. This value is known as the *statistical power* of the test, the probability of rejecting the null hypothesis when it is false. We consider the issue of statistical power in more detail later in the chapter.

According to the traditional null hypothesis significance testing view,[4] if the null hypothesis can be rejected as being unlikely (typically $p < .05$), then we can conclude that an effect of the treatment exists. $\bar{Y}_t - \bar{Y}_c$ provides an unbiased estimate of the causal effect of the treatment. Once again, this estimate is accurate on average; however, there is no guarantee that the estimate will be accurate in any specific experiment – the value may be too high or too low. These considerations highlight the value of replications and meta-analyses (see e.g., Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Johnson & Eagly, this volume, Ch. 19) in establishing more accurate and more certain estimates of causal effects.

## FIELD EXPERIMENTS

Both basic and applied social psychological experiments may be carried out in field settings. The defining characteristics of the field experiment closely follow those of the laboratory experiment in social psychology (Aronson, Wilson, & Brewer, 1998; Smith, this volume Ch. 12) – random assignment, manipulation of the treatment conditions, and measurement of the dependent variable. However, additional issues arise because of differences between the field and laboratory

[3] Researchers typically only see the causal effect estimate from their own single experiment. They have worked hard designing the experiment, recruiting and running the participants, and analyzing the data and consequently have great confidence in the results of their single experiment. When another researcher fails to find the same result, it is easy to attribute his lack of findings to methodological problems (the "crap research" hypothesis; Hunter, 1996). However, Hunter argued that meta-analyses of several research areas have suggested that methodological quality accounts for relatively little of the variability in the causal effects. Rather, simple sampling variability of the type illustrated here appears to be the main source of the variability in estimates of causal effects (Hunter, 1996).

[4] At the time of this writing, intense reconsideration of hypothesis testing procedures is taking place among methodologists (see Harlow, Mulaik, & Steiger, 1997; Wilkinson, L. and the Task Force on Statistical Inference, 1999). Our belief is that traditional null hypothesis testing procedures will likely be revised or replaced by an alternative procedure during the next few years. One major alternative is that researchers will be asked to report confidence intervals for all effect estimates. Confidence intervals that do not include 0 lead to the same conclusions as null hypothesis significance tests with respect to ruling out chance as a plausible interpretation of the results. But, confidence intervals also provide information about the size of the treatment. Standardized and unstandardized measures of the size of treatment effects are likely to become increasingly important in many areas of psychology in the future (Cohen, Cohen , Aiken, & West, in press).

contexts. Below we first briefly present two simplified examples of field experiments, then identify some of the new issues that arise as well as possible solutions for those issues.

## Illustrations of Field Experiments

Some field experiments test hypotheses derived from basic social psychological theory. For example, Cialdini, Reno, and Kallgren (1990) conducted a series of field experiments to examine the impact of norms on littering. In one study on the impact of descriptive norms (what most others do), solitary dormitory residents retrieving their mail (which contained a flier) were confronted with an environment containing either no litter, a single conspicuous piece of litter, or many pieces of litter. More residents in the heavily littered environment condition discarded the flier than in the other two conditions. However, fewer residents in the single piece of litter condition littered than in the clean environment condition. Cialdini et al. argued that the single piece of litter invoked the descriptive norm against littering.

Other field experiments test hypotheses from applied social psychology. For example, Evans et al. (1981) evaluated the effectiveness of a program to deter smoking among junior high school students in Houston. The program built on social psychological theory and research (e.g., Bandura, 1977; McGuire, 1964). The program included three major components: (a) social pressures created by peers, family, and the media to smoke; (b) the immediate negative physiological effects of smoking; and (c) methods of resisting pressures to smoke. The program was intended to be of 3 years duration. Entire schools were assigned to program (treatment) and no treatment (control) conditions. Although the great majority of students participated, informed consent requirements resulted in a sample of self-selected volunteers. Complex methodological issues arose in this experiment: How to assign the small number of units (schools) to treatment conditions, how to analyze data from students in the treatment condition who received from 1 to 3 years of the program because of school transfers, and how to address the loss of participants at the posttest measurement. At the end of the 3-year experiment, students in the smoking prevention program reported lower intentions to start smoking and lower rates of cigarette smoking than did control students. A number of experiments evaluating other smoking prevention programs derived from Evans et al.'s original work have shown generally positive results in preventing or delaying the onset of smoking in preteenagers (Flay, 1986).

## Problems of Randomized Experiments and Their Remedies

**THE STABLE-UNIT-TREATMENT-VALUE-ASSUMPTION (SUTVA).** Randomization yields unbiased estimates of the causal effect given that the SUTVA is met. SUTVA has two parts. First, the assignment mechanism, here randomization, should not affect the participants' response to the experimental or control treatment. Although we normally consider randomization to be an inert process with respect to the participant's response, consider the following thought experiment. A student volunteers for an exciting new graduate training program in which she is handsomely funded each summer of her graduate career to work with a different professor of her choice in her speciality area at any university in the United States. She is informed that she has been chosen for this program on the basis of (i) her extraordinary merit, or (ii) random assignment. Would the participant's response (e.g., measured achievement in research) to the same program be identical in each case? At least some social psychological theorizing (e.g., Weiner, 1974) would predict that a student who believed good luck was responsible for her program participation might show lower achievement than the same student who believed her own high levels of ability and effort were responsible for her program participation.

Second, the response of the participant should not be affected by the treatments other units receive. Knowledge of other treatments may change the participants' level of motivation, particularly for those in the control group, a problem Higginbotham, West, and Forsyth (1988) termed "atypical reactions of control participants." For example, participants assigned to the control group in a randomized experiment of a promising new treatment for a serious disease may give up their normal health-maintaining activities, leading to poorer outcomes in the control group than normal (resentful demoralization). Alternatively, participants assigned to a well-liked standard (comparison) program may try harder if the existence of the standard program may be in jeopardy as a result of a positive outcome for a new, experimental program (compensatory rivalry). Such atypical responses of control participants can lead to overestimates or underestimates of causal effects, respectively. The participant's response can also

be affected in other ways when the response involves a comparative evaluation. For example, the causal effect of a job training program relative to a no training control group may be overestimated in a small community if most of the limited pool of available jobs are then taken by trainees, so that there are few, if any, of the usual number of possible positions available for control group members.

Researchers have developed a variety of strategies to increase the likelihood that SUTVA is met, of which the following three are the most commonly used.

1. If participants are not aware of the random assignment and are only aware of the nature of the experimental condition in which they participate, the likelihood that SUTVA will be met is greatly increased. Such conditions can typically be achieved in laboratory settings[5] and some field settings. In particular, geographical or temporal isolation of the participants in the treatment and control conditions can often help minimize these problems in field settings. However, a caution is in order here: Informed consent procedures sometimes mandate informing participants that they will be randomly assigned to one of several treatment conditions whose nature is outlined in the consent form. Such procedures may potentially lead to violations of SUTVA.
2. Successful masking (blinding) procedures in which the participants (and ideally the experimenter as well) are kept unaware of their treatment condition also increase the likelihood that SUTVA will be satisfied. Such procedures are commonly used in drug research in which participants are not informed whether they are receiving an active drug or an inert placebo. However, masking procedures often cannot be successfully applied in real world experiments. For example, consider an experiment (Mallar & Thornton, 1978) in which released prisoners are given transitional financial aid for 6 months (treatment) or no financial aid, and the researchers recorded whether they return to jail for property crimes during the following year. Keeping released prisoners unaware that they have received financial assistance would eliminate any positive effects this treatment could be expected to have. Further, treatment masking is not always successful: Participants sometimes see through even the best masking procedures on the basis of outcomes they experience early in the experiment (Meier, 1991). Even in the most carefully conducted drug trials, participants sometimes can identify the medication they have been assigned based on factors such as its taste, early positive effects, and early side effects. Participants who have strong positive (or negative) expectations about the effects of that drug may then show increased positive (or negative) responses to the drug.
3. The specific treatments do not represent outcomes or opportunities that are important to participants. Participants assigned to a 6-person simulated jury in a laboratory experiment are unlikely to change their responses on the basis of their knowledge that other participants have been assigned to a 12-person simulated jury. In contrast, released prisoners in the control condition may well be angered or demoralized if they learn that other former prisoners in the treatment condition receive 6 months of transitional aid to cover basic living expenses. An alternative strategy is to offer control (comparison) participants an equally attractive program that is not expected to affect the response. Bryan, Aiken and West (1996) presented an example of this design, comparing the effectiveness of an STD-(sexually transmitted disease) prevention program with an equally attractive stress-reduction comparison program. The stress-reduction comparison program did not include any content that focused on increasing condom use and was thus not expected to influence the response of interest, condom use.

**BREAKDOWN OF RANDOMIZATION.** For the randomized experiment to yield an unbiased estimate of the causal effect, $\bar{Y}_t - \bar{Y}_c$, random assignment of participants to treatment and comparison conditions must, in fact, be properly carried out. Studies of large-scale randomized field experiments, particularly those conducted at multiple sites, suggest that full or partial breakdowns of randomization occur with some frequency (Boruch, McSweeny, & Soderstrom, 1978; Conner, 1977). Problems tend to occur more frequently when the individuals responsible for delivering the treatment, such as school or medical personnel, are allowed to carry out the assignment of participants to treatment conditions and when monitoring of the

[5] SUTVA may be violated in laboratory experiments through prior communication of information about the experiment among potential participants. The little research to date (e.g., Aronson, 1966) suggests that this may be a minor problem, at least among unacquainted participants, given adequate debriefing following the experiment. Potentially more problematic, but little researched, is the prevalence and effects of communication through informal social networks of acquaintances who may seek information prior to signing up for a particular experiment.

maintenance of the treatment assignment is poor. For example, Kopans (1994) reviewed the large recent Canadian experiment on the effectiveness of mammography screening for reducing deaths from breast cancer. He presented data suggesting that women in the mammography group had a substantially higher cancer risk *at pretest* than women in the no mammography screening group. It seems likely that some of the physicians involved in the trial saw to it that their patients with family histories of breast cancer or prior episodes of breast-related disease were assigned to the screening group.

One way to address this problem is through careful monitoring of the randomization process as well as careful monitoring of the treatment(s) each participant actually receives following randomization (Braucht & Reichardt, 1993). For example, students (or their parents) in school-based experiments are sometimes able to agitate successfully to change from a control to a treatment class during the randomization process itself or during the school year following randomization. Careful monitoring can help minimize this problem and can also allow formal assessment of the magnitude of the problem. If there is a systematic movement of children between treatment conditions (e.g., the brighter children are moved to the treatment group), the estimate of the causal effect of treatment will potentially be biased. Attempts to correct for such bias need to be undertaken.

An alternative strategy to minimize breakdowns of randomization is to use units that are temporally or geographically isolated in the experiment. For example, randomization breakdowns in school-based experiments, particularly those which occur postassignment, are far more likely when different treatments are given to different classrooms (low isolation of units) than when different treatments are given to different schools (high isolation of units).

**GROUP ADMINISTRATION OF TREATMENT.** Often in field research interventions are offered to groups of participants. For example, Evans et al. (1981) delivered their smoking prevention intervention to intact school classrooms; Aiken, West, Woodward, Reno, and Reynolds (1994) delivered a program encouraging compliance with American Cancer Society guidelines for regular screening mammograms to women's groups in the community; and Vinokur, Price, and Caplan (1991) recruited unemployed individuals to participate in a job-seeking skills training program that was delivered in a group format. Group administration of treatments, whether in the laboratory or the field, leads to statistical and conceptual issues that need to be considered.

When treatments are delivered to groups, the entire group is assigned to either the treatment or control condition. Thus, randomization occurs at the level of the group and not the individual participant. The statistical outcome of this procedure is that the responses of the members of each treatment group may no longer be independent. As an illustration, consider the Evans et al. (1981) smoking prevention experiment. The responses of 2 children randomly chosen from a single classroom would be expected to be more similar than the responses of 2 children randomly chosen from different classrooms (Kashy & Kenny, this volume, Ch. 17). Although nonindependence has no impact on the causal effect estimate, $\bar{Y}_t - \bar{Y}_c$, it does lead to estimates of the standard error of this effect that are typically too small.[6] The magnitude of this problem increases as the amount of dependence (measured by the intraclass correlation) and the size of groups to which treatment is delivered increase. For example, Barcikowski (1981) showed that, even with relatively low levels of dependency in groups (intraclass correlation = .05), when groups were of a size typical of grade school classroom studies ($n = 25$ per class), the Type 1 error rate (rejecting the null hypothesis when in fact it is true) was in fact .19 rather than the stated value of .05.

Following Fisher (1935), researchers traditionally "solved" this problem by aggregating their data to the level of the group (e.g., average response of each classroom). The unit of analysis should match the unit of assignment – "analyze them as you've randomized them" (Fisher, as cited in Boruch, 1997, p. 195). However, this solution is often not fully satisfactory because such analyses can be generalized only to a population of groups, not to a population of individuals, which is typically the researcher's interest. Over the past decade, new statistical procedures termed "hierarchical linear (random coefficient, multilevel) models" have been developed. These models simultaneously provide an estimate of the causal effect at the group level as well as individual level analyses that appropriately correct standard errors for the degree of nonindependence within groups. Introductions to these models are presented in Bryk and Raudenbush (1992), Hedecker, Gibbons, and Flay, (1994), and Kreft and DeLeeuw (1998).

[6] In randomized field experiments, participants are nearly always more similar within than between groups. In other situations involving within-subject designs or other forms of dependency, the direction of bias may change (see Kenny & Judd, 1986).

Given the proper use of hierarchical linear models to analyze the data, the random assignment of groups to treatments (when participants are also randomly assigned to the individual groups) rules out traditional sources of confounding, known as ecological bias, associated with aggregation of units (Greenland & Morgenstern, 1989; Robinson, 1950). However, a conceptual issue remains. Rubin's causal model emphasizes that causal effects represent the comparison of one well-articulated treatment with another well-articulated treatment. In comparisons among treatments delivered to individuals in group settings, the articulation of the treatment becomes murkier as treatment now includes the other individuals in the setting and all of the activities of members within the group. Cronbach (1976) and Burstein (1980) discussed many of the issues and opportunities associated with multilevel designs, although their recommendations of analytic strategies have been superseded by the hierarchical linear modeling approaches noted previously. Draper (1995), Holland (1989), Burstein (1985), and Burstein, Kim, and Delandshire (1989) considered many of the inferential issues associated with hierarchical linear modeling.

**TREATMENT NONCOMPLIANCE.** When participants are randomly assigned to treatment and control conditions in field experiments, not all participants may actually get the treatment. A portion of the participants randomly assigned to programs designed to increase exercise, improve nutrition, or improve job-seeking skills simply do not show up for program sessions, and so do not receive the treatment. These participants are referred to as treatment noncompliers. Practical methods exist for minimizing this problem, notably making the program attractive to participants, removing barriers to program attendance (e.g., providing transportation or child care), giving participants incentives for program attendance, and only including those participants who are willing to participate in both the treatment and control programs in the randomization (Cook & Campbell, 1979). Despite these efforts, some participants assigned to treatment condition never receive any treatment. We assume in this section that the researcher was able to measure the dependent variable on all participants, including those who do not receive treatment.

Three statistical approaches have been taken in response to this problem. The first, known as the *intention to treat* analysis (Lee, Ellenberg, Hirtz, & Nelson, 1991) is to compare the mean response of all participants assigned to the treatment condition (regardless of whether they received treatment) with the mean response of all participants assigned to the control condition. This analysis typically yields conservative estimates of the causal effect and requires no assumptions beyond those required for the randomized experiment. The second is to throw out all participants assigned to the treatment group who do not in fact receive treatment. Such a comparison will yield a biased estimate of the causal effect (with the direction of bias being unknown) unless the stringent assumption can be made that the participants who drop out of the treatment condition represent a random sample of the participants in that condition.[7] The third, known as the complier average causal effect (CACE), uses ideas from econometrics and missing data theory (Little & Rubin, 1987) to create an unbiased estimate of the treatment effect for participants who actually receive the treatment (Angrist et al., 1996; Little & Yau, 1998; see also Bloom, 1984). Both the first and the third approaches produce meaningful estimates of treatment effects; however they answer different questions (see West & Sagarin, 2000). The intention to treat analysis estimates the causal effect in the entire sample, whereas CACE estimates the causal effect only for those participants who actually receive the treatment.

To understand these three approaches, let us consider the data presented in Table 3.4. Table 3.4 uses the same data as Table 3.1 with some added features. First, in assignment 4 ($a_4$) Participants 1–16 who are assigned to the control group are assumed to be perfectly matched with Participants 17–32 who are assigned to the treatment group. In this assignment, Participant 1 is identical to Participant 17, Participant 2 is identical to Participant 18, and so on, so that we do not need to consider the effects of sampling error. Second, we have indicated a systematic pattern of noncompliance. In column $c_1$, participants with the lowest scores prior to treatment do not comply with the treatment. These participants are termed "never takers" in Angrist et al.'s (1996) model.

The use of the RCM, which emphasizes the comparison of the same unit receiving the treatment and the control conditions, highlights an easily overlooked point. Participants 1–5 in the control group are identical to the never takers in the treatment group (Participants 17–21) in Table 3.4. They are participants who would

[7] The approach of throwing out participants has been the standard procedure in laboratory experiments in social psychology when technical problems arise or when participants are suspicious or uncooperative. The possibility that this procedure introduces potential bias should always be considered.

**TABLE 3.4. Illustration of Effects of Treatment Noncompliance**

| Participant | $a_4$ | $c_1$ | $Y_c$ | $Y_t$ |
|---|---|---|---|---|
| 1 | 0 | 0 | 1* | 1.5 |
| 2 | 0 | 0 | 2* | 2.5 |
| 3 | 0 | 0 | 2* | 2.5 |
| 4 | 0 | 0 | 2* | 2.5 |
| 5 | 0 | 0 | 2* | 2.5 |
| 6 | 0 | 1 | 3* | 3.5 |
| 7 | 0 | 1 | 3* | 3.5 |
| 8 | 0 | 1 | 3* | 3.5 |
| 9 | 0 | 1 | 3* | 3.5 |
| 10 | 0 | 1 | 3* | 3.5 |
| 11 | 0 | 1 | 3* | 3.5 |
| 12 | 0 | 1 | 4* | 4.5 |
| 13 | 0 | 1 | 4* | 4.5 |
| 14 | 0 | 1 | 4* | 4.5 |
| 15 | 0 | 1 | 4* | 4.5 |
| 16 | 0 | 1 | 5* | 5.5 |
| 17 | 1 | 0 | 1* | 1.5 |
| 18 | 1 | 0 | 2* | 2.5 |
| 19 | 1 | 0 | 2* | 2.5 |
| 20 | 1 | 0 | 2* | 2.5 |
| 21 | 1 | 0 | 2* | 2.5 |
| 22 | 1 | 1 | 3 | 3.5* |
| 23 | 1 | 1 | 3 | 3.5* |
| 24 | 1 | 1 | 3 | 3.5* |
| 25 | 1 | 1 | 3 | 3.5* |
| 26 | 1 | 1 | 3 | 3.5* |
| 27 | 1 | 1 | 3 | 3.5* |
| 28 | 1 | 1 | 4 | 4.5* |
| 29 | 1 | 1 | 4 | 4.5* |
| 30 | 1 | 1 | 4 | 4.5* |
| 31 | 1 | 1 | 4 | 4.5* |
| 32 | 1 | 1 | 5 | 5.5* |

*Note.* The column labeled $Y_c$ contains the true response of each participant in the control condition. The column labeled $Y_t$ contains the true response of each participant in the treatment condition. $a_4$ represents the assignment of the first 16 participants to the control group and the second 16 participants to the treatment group. $c_1 = 1$ means participant follows the treatment or control condition as assigned. $c_1 = 0$ means participant is a never taker and does not comply when in the treatment condition. The starred value of $Y$ is the value actually observed for each participant. As before, the true causal effect, $Y_t - Y_c$, is 0.5.

not have complied with the treatment if they had been assigned to the treatment group. All standard analyses that throw out noncompliers fail to take into account this group of participants who would fail to comply if they were given the opportunity. As will be illustrated below, the failure to consider this group of participants potentially yields biased estimates of treatment effects.

True diff is .5

Applying these three statistical approaches to the present example, the intention to treat analysis compares the observed data (indicated with an asterisk in Table 3.4) for participants assigned to the treatment with the observed data for participants assigned to the control group. The mean for the control group is as before, $\bar{Y}_c = 3.0$. However, in the treatment group Participants 17–21 did not comply and thus received no benefit from treatment. Consequently, the treatment group mean is correspondingly reduced, $\bar{Y}_t = 3.344$. The causal effect estimate, $\bar{Y}_t - Y_c$, is 0.344 – more than a 30% reduction in the effect size from the true value for compliers of 0.50.

Intent to treat diff is .34

Following the second approach, Participants 17–21 are eliminated from the analysis and the mean for the treatment group is $\sum_{i=22}^{i=32} Y_i/11 = 4.045$. Thus, the causal effect estimate is $\bar{Y}_t - \bar{Y}_c = 4.045 - 3.000 = 1.045$, which in this case is considerably larger than the true value of 0.5. Finally, using the CACE approach (Little & Yau, 1998), we eliminate the never takers from both the treatment and control groups. We find that

typical analysis diff is 1.045

$$\bar{Y}_t = \frac{\sum_{i=22}^{i=32} Y_i}{11} = 4.045, \quad \bar{Y}_c = \frac{\sum_{i=6}^{i=16} Y_i}{11} = 3.545,$$

so that the estimate of the causal effect, $\bar{Y}_t - \bar{Y}_c$, is 0.5, which is equal to the true effect for compliers.

The conceptual problem with CACE is that we cannot identify which participants in the control group would comply if they were in the treatment group. However, given a randomized experiment and SUTVA, an unbiased estimate of CACE can be calculated. In the simplest case (Bloom, 1984), in which we assume the treatment effect is constant for all participants, the causal effect estimate, $\bar{Y}_t - \bar{Y}_c$, from the intention to treat analysis (.344) can be adjusted by the inverse of the proportion of compliers in the treatment group $(11/16)^{-1}$. In the present example, this effect can be calculated as CACE = (.344)(16/11) = 0.5. Angrist et al. (1996) and Little and Yau (1998) provided more advanced statistical discussions of the CACE approach. Vinokur, Price, and Caplan (1991) provided an empirical illustration.

CACE diff is .5

The CACE approach makes several assumptions. For field experiments in social psychology, the most important of the assumptions is that the response of a never taker participant who does not comply with the treatment will be identical to the response of the same

participant in the control group. This assumption implies that the treatment received by noncompliers in the treatment group must be identical to the treatment received by participants in the control group. It will only be met in designs in which the control group represents "no treatment" or in which the control group represents a base treatment ($t_{base}$, which everyone receives) to which one or more additional components are added in the treatment group ($t_{base} + t_{additional}$; the constructive research strategy; see West & Aiken, 1997). Designs in which an alternative treatment is used as the comparison group will violate this assumption.

The assumption will also be violated in experiments in which participants partially comply with treatment (e.g., they attend 5 of 20 required treatment sessions). Attempts to adjust intention to treat causal effect estimates based on careful measures of degree of compliance of participants in *both* the treatment and control groups have thus far not been fully satisfactory. They require either an excellent model of the determinants of compliance in both the treatment and control groups or the use of a successful double-masking (blinding) procedure, in which neither the participant nor the experimenter is aware of the participant's treatment condition. Holland (1988) presented an extensive discussion of this problem in terms of the RCM. Efron and Feldman (1991) presented one of the best of the current procedures for addressing partial treatment compliance. The discussions following both the Holland and the Efron and Feldman articles clearly articulate the complex issues and limitations associated with the use of such procedures.

**PARTICIPANT LOSS AT POSTTEST MEASUREMENT.** The approaches for addressing treatment noncompliance just discussed assumed that posttest measurements were available for all participants regardless of whether they complied. A second problem that occurs in many randomized field experiments is that some participants in both the treatment and comparison groups cannot be remeasured at posttest. This problem, known as participant *attrition*, is a major potential source of bias in the estimation of causal effects.

Attrition can often be minimized by careful attention during the planning of the experiment. Securing the addresses and telephone numbers of the participants, their close friends or relatives, and their employer or school greatly aids in locating dropouts (attriting participants). Keeping in touch with both treatment and control participants through periodic mailings or telephone calls and providing incentives for continued participation can also help minimize participant loss. A considerable body of specialized techniques for tracking, locating, and contacting participants now exists in many research areas (Ribisl, Walton, Mowbray, Luke, Davidson, & Bootsmiller, 1996). Nonetheless, even given the use of careful procedures, attrition still typically does occur. For example, Biglan et al. (1991), in their review of longitudinal studies of substance abuse prevention, reported attrition rates ranging from 5% to 66% (mean = approximately 25%). Furthermore, dropouts typically report greater substance use at the initial measurement. Such findings suggest that estimates of treatment effects on the outcomes of interest may be biased if attrition is not addressed.

Researchers have some ability to estimate the likely effects of attrition on their results – but only if pretest measures that are expected to be related to the response measures of interest have been collected. Jurs and Glass (1971; see also Cook & Campbell, 1979) have outlined a two-step strategy to detect possible bias introduced by differential attrition between the treatment and comparison groups.

1. The first step is to compare the percentage of participants who drop out in the treatment and comparison groups. If these values differ, then the differential attrition (participant loss) rates may be interpreted as a treatment effect. However, the interpretation of all measured response variables becomes more problematic (see Sackett & Gent, 1979).
2. The second step is to conduct a series of 2 (treatment group: t vs. c) × 2 (completer vs. attriter) ANOVAs on each of the pretest measures. The goal is to identify all possible background characteristics (e.g., attitudes, behaviors, personality, demographics) on which participants may differ. A main effect for treatment group indicates a possible failure of the randomization to properly equate the groups prior to treatment. A main effect for attriter versus completer indicates that the characteristics of the attriters (dropouts) differ from those of the completers, suggesting that the findings of the experiment cannot be generalized to the full population of interest (external validity threat). Of most concern, a Treatment Group × Attrition Status interaction indicates that the characteristics of participants who dropped out differed in the treatment and control conditions, indicating potential bias in the estimate of the causal effect (internal validity threat). When problems of differential attrition are identified, researchers should explore possible adjustments of the treatment effect through missing data imputation techniques (Little & Rubin, 1987; Little & Schenker, 1995), through attempts to understand

and model the effects of attrition on the response (e.g., Arbuckle, 1996; Muthén, Kaplan, & Hollis, 1987), or through attempts to estimate reasonable maximum and minimum values that bracket the treatment effect (Shadish, Hu, Glaser, Knonacki, & Wong, 1998; West & Sagarin, 2000). Comparison of alternative adjustments that make different assumptions to see if they yield similar results is particularly worthwhile.

There are two major problems in the use of the Jurs and Glass (1971) technique just outlined. First, differential attrition may be related to participant characteristics that were not measured at pretest. If these unobserved characteristics differ between the treatment and control groups and are related to the responses of interest, the causal effect estimate may be misleading. For example, it is possible that attriters from smoking prevention programs may have peers or parents who smoke and who put pressure on them to drop out of the program. In contrast, children in the control group with parents or peers who smoke would not experience similar pressure and would be less likely to drop out. If measures of peer and parent smoking are not collected at pretest, this potential source of differential attrition could not be detected. Second, the Jurs and Glass technique uses statistical hypothesis tests to identify variables that are related or unrelated to attrition status. Concluding that a variable is unrelated to attrition status requires the researcher to accept the null hypothesis that there is no Treatment × Attrition Status interaction. In fact, some of the individual pretest variables, or combinations of these pretest variables, may represent important sources of differential attrition even though the statistical test fails to reach conventional levels of significance. Some researchers (e.g., Hansen, Collins, Malotte, Johnson, & Fielding, 1985) have called for using less conservative levels of significance (e.g., $\alpha = .25$) for all tests so that the corresponding Type II error rates will be reduced. Others (Tebes, Snow, & Arthur, 1992) have proposed choosing the Type I error rate ($\alpha$) based on an expected or calculated effect size so that the ratio of Type II to Type I errors will approximate the ratio of 4:1 (viz. $\beta = .20$; $\alpha = .05$) recommended by Cohen (1988). The bottom line is that rejection of the null hypothesis is used as a method of screening potentially important from unimportant sources of differential attrition. The basic goal is to identify all potential sources of differential attrition that should be retained for further investigation. Even if a large number of tests are performed, they should not be corrected for alpha inflation (experimentwise error rate). Such correction defeats the basic goal of the procedure.

## OTHER ISSUES IN EXPERIMENTS

### Statistical Power

For a randomized trial to be worth doing, it must have adequate statistical power to detect differences between the treatment and control conditions. Yet, existing reviews of the statistical power of tests in major psychology journals, including the *Journal of Personality and Social Psychology* (e.g., Rossi, 1990), have suggested that many researchers continue to conduct statistical tests with inadequate power. Following Cohen (1988), differences between the treatment and control groups of .80, .50, and .20 standard deviation units, ($d$), are defined as large, moderate, and small effect sizes, and .80 is defined as adequate statistical power. In a randomized experiment with an equal number of participants in the treatment and control groups, 52 total participants ($n = 26$ per condition) would be needed to detect a large effect, 126 participants would be needed to detect a moderate effect, and 786 participants would be required to detect a small effect at alpha = .05 and power = .80. Our earlier example illustrated this issue: With 32 total participants, we were able to detect a moderate effect size ($d = 0.5$) only about 30% of the time. Given an estimate of the likely effect size based either on prior research in the area (especially estimates from meta-analyses) or on a normative basis (e.g., $d = .50$ following Cohen, 1988), statistical power can now be easily computed in advance of conducting an experiment with user-friendly software (e.g., Borenstein, Cohen, & Rothstein, 1997).

In the design of experiments, several methods can be used to increase power. These methods are extensively discussed in Higginbotham, West, and Forsyth (1988, chapter 2); Hansen and Collins (1994); Lipsey (1997); and Dennis, Lennox, and Williams (1997). The most obvious method is to increase the sample size. However, practical issues such as cost, available resources, or the limited availability of certain special participant populations (e.g., bereaved children; intensive care nurses) often restrict this option, even in large metropolitan areas. Other methods include using stronger treatments, maintaining the integrity of the treatment implementation, using more reliable measures, maximizing participant uniformity, minimizing participant attrition, and adding covariates measured at pretest that are related to the dependent variable. Also of interest are a variety of rarely used techniques that equate participants on one or more covariates that are known to be important predictors of the response prior to random assignment to treatment and control conditions. Such procedures are particularly valuable

when only a small number of units are available for randomization.

Equating participants at pretest on a single variable is straightforward. Typically, participants are simply ranked on the background variable and then grouped into pairs (the 2 highest, the next 2 highest, and so on, down to the 2 lowest scores on the pretest measure). Then, 1 participant from each pair is randomly assigned to the treatment and the other to the control condition. A matched-pairs $t$-test is used to compare the response of the treatment and control groups. If the basis for matching participants is highly related to the response variable, these procedures can lead to large increases in statistical power. For example, Student (1931, the nom de plume of W. S. Gosset, a statistician at the Guinness brewing company) critiqued the Lanarkshire milk experiment, an early experiment comparing the gains in height and weight of 5,000 children given pasteurized milk with those of 5,000 children given raw milk. He noted that identifying 50 pairs of identical twins and then randomly assigning 1 twin from each pair to the pasteurized milk condition and 1 twin to the raw milk condition would yield the same statistical power as the original experiment with 10,000 children.

With several variables, these procedures become more complex. To illustrate one method of matching on several variables, imagine that the 32 participants in our earlier example had been measured on several important background variables at pretest. We noted earlier that there were over 600 million different possible randomizations of the 32 participants into two equal groups ($n_1 = n_2 = 16$). We could have a computer program generate a large number (e.g., 1,500) of different randomizations. A multivariate measure (e.g., Hotelling's $T^2$) of the difference between the two groups on the pretest measures would be calculated for each of the 1,500 randomizations. We could then rank order the randomizations according to their success in equating the treatment and control groups on the pretest measures on the basis of the multivariate measure. We would then randomly select one of the best randomizations (e.g., from the 50 with the lowest value of the Hotelling's $T^2$) to use in our experiment. Such procedures assure that our two groups are well-equated at pretest, minimizing sampling error to the extent the pretest measures are related to the response.[8]

[8] Analysis of covariance or blocking on the pretest score can also be used to increase statistical power. These techniques are more sensitive to violations of assumptions (e.g., curvilinear relationship between pretest measure and response) than matching and generally perform less well in small samples. Maxwell and Delaney (1990, chapter 9) presented a thorough discussion and comparison of matching and blocking techniques in experiments.

### Generalization of Causal Relationships

Our presentation of the RCM has focused primarily on what Cook and Campbell (1979) have termed *internal validity*: Something about the particular treatment caused the specific observed response in this particular experiment. The RCM is focused on the estimation of the causal effect; it is largely mute as to what the active causal agent(s) of the treatment might be or how one might go about determining them. Campbell (1986) also emphasized the limited causal understanding provided by internal validity, even suggesting that internal validity be relabeled as "local molar (pragmatic, atheoretical) causal validity" (p. 69).

In contrast, social psychological researchers are rarely interested in limiting their causal statements to a specific treatment implementation, delivered in a specific setting, to a specific sample of participants, and assessed with a specific measure. Methodologists (Campbell, 1957; Campbell & Stanley, 1966; Cook, 1993; Cook & Campbell, 1979; Cronbach, 1982; Shadish et al., in press; see also Brewer, this volume, Ch. 1) have articulated principles for understanding and generalizing the causal effect obtained in a single experiment. Cronbach (1982) developed a model with a focus on generalization with respect to four dimensions: units (typically participants), treatments, observations (measures or responses), and settings. He labeled the specific values of these dimensions that characterize a specific experiment as (lower case) **u**nits, **t**reatments, **o**bservations, and **s**ettings (utos). He labeled the target values of the dimensions that characterize the classes to which the results of the experiment can be generalized as (upper case) **U**nits, **T**reatments, **O**bservations, and **S**ettings (UTOS). In Cook and Campbell's (1979, chapter 2) analysis of validity, generalization to Treatments and Observations represent two dimensions of the issue of construct validity, and generalization to Units and Settings (and Times) represent the dimensions of the issue of external validity. Cronbach (1982) also described another type of generalization, utos to U*T*O*S*. Here, the researcher desires to generalize to new populations of Units, Treatments, Observations, and Settings (U*T*O*S*) that were not studied in the original experiment.

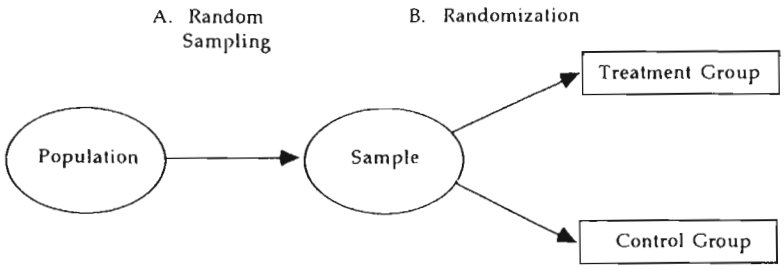



**Figure 3.2.** The formal statistical model for generalization.

*Note.* The purpose of step B is to provide unbiased estimates of the causal effect of the treatment. The purpose of step A is to permit generalization of the results obtained in the sample to a defined population of participants.

Let us apply the Cronbach model to the Evans et al. (1981) experiment described earlier. In this experiment, the units were children in specific school classrooms in Houston, the treatment was this specific implementation of the social influences smoking prevention program, the observations were children's reports of smoking, and the setting was the classroom. The UTOS to which Evans et al. presumably wished to generalize were all school children of specific ages, the social influences smoking prevention program, cigarette smoking, and school classrooms. As an applied experiment, the goal is to generalize to specific participant populations and to a specific school setting. The treatment is also conceptually quite circumscribed as is the observation of interest, smoking.

## Strategies for Generalization

**STATISTICAL STRATEGY: RANDOM SAMPLING.** The only formal statistical basis for generalization is through the use of random sampling from well-defined populations. Surveys using national probability samples assess the attitudes of representative samples of adults; other surveys may assess the attitudes of a more focused sample, such as hospital nurses in a section of Denver. In each case, a defined population is enumerated from which a random sample is collected, yielding estimates of the attitudes of the population that are accurate to within a specified level of sampling error.[9]

Figure 3.2 presents the randomized experiment in the context of this formal sampling model. Stage A represents random sampling from a defined population; the purpose of this stage is to assure generalization to a defined population of Units (participants) as is discussed below. Stage B represents random assignment of the units in the sample to treatment and control conditions; as discussed previously, the purpose is to achieve unbiased estimates of the causal effect in the sample. The combination of Stages A and B formally permits generalization of an unbiased causal effect of the specific treatment conditions, studied in the context of a specific experimental setting, using the specific dependent measures to the full population of Units (participants). Note that generalization to Treatments, Observations, and Settings is not formally addressed by this model.

This formal sampling model is routinely recommended by statisticians and epidemiologists (e.g., Draper, 1995; Kish, 1987) as the ideal model for experimental research. Unfortunately, it is extraordinarily difficult to implement in practice. With respect to Units (participant populations), many cannot be precisely enumerated (e.g., children whose parents are alcoholics; Chassin, Barrera, Bech, & Kossak-Fuller, 1992). Even when the researcher can precisely enumerate the participant population (e.g., using official court records to enumerate recently divorced individuals), there is no guarantee that such participants can actually be located. Further, even if they are located, participants may refuse to be randomized or refuse to participate in a particular experiment, despite being offered large incentives for their participation.

A few recent experiments have approximated the ideal model of statisticians despite the difficulties. Randomized experiments have been conducted within national or local probability surveys (e.g., investigating question context; Schwarz & Hippler, 1995). Randomized experiments have compared treatment versus control programs using random samples of specific populations of individuals in the community (e.g., job seekers selected from state unemployment lines;

[9] Social psychologists interested in basic research have traditionally focused primarily on whether theoretically predicted effects exist. However, recent criticisms of traditional null hypothesis significance testing (e.g., Cohen, 1994; Harlow, Mulaik, & Steiger, 1997) and the increased prominence of meta-analysis have greatly increased the focus of both basic and applied psychologists on the size of treatment effects. This shift in focus is likely to lead to a stronger focus on issues of generalization in basic as well as in applied research in the future.

Vinokur et al., 1991; Wolchik et al., in press). Such experiments routinely include heroic efforts to study nonparticipants in the experiment in order to understand the probable limits, if any, on the generalization of their findings to the full participant population of interest. Nonetheless, even such extraordinary efforts only address the generalization of units; they only take us from u to U. When our interest turns to the generalization of findings to a class (or population) of Treatments, Observations, and Settings, we almost never have any strong basis for defining a population of interest. In short, in nearly all basic and applied social psychological research we have to turn to extra-statistical methods of enhancing causal generalization.

**EXTRA-STATISTICAL APPROACHES: COOK'S FIVE PRINCIPLES.** Cook (1993) synthesized earlier ideas about causal generalization and articulated five general principles. These principles may be applied to strengthen causal generalization with respect to Units, Treatments, Observations, and Settings.

PROXIMAL SIMILARITY. The specific units, treatments, observations, and settings should include most of the components of the construct or population, particularly those that are judged to be prototypical. A researcher wishing to generalize to a population of nurses (e.g., in metropolitan Denver) should choose nurses from this area in his sample. The sample should include the various modal types of nurses (e.g., LPN, RN). To generalize to settings, the modal settings in which nurses work (e.g., hospital, home-care) should be identified, and nurses should be sampled from each. With respect to constructs, the researcher should design a treatment and either design or select a measurement instrument that includes most of the important components specified by the theory.

HETEROGENEOUS IRRELEVANCIES. The units, treatments, observations, and settings we use in our experiments are specific instances chosen to represent the population or the construct. They will typically underrepresent certain features of the target Units, Treatments, Observations, and Settings. They will typically also include other extraneous features that are not part of the target of generalization. For example, nearly all research on attitudes uses paper-and-pencil measurement techniques, yet paper-and-pencil measurement is not part of the definition of attitudes (see Sears, 1986, and Houts, Cook, & Shadish, 1986, for other examples of these issues). Following the principle of heterogenous irrelevancies calls for the use of multiple instances in our research that are heterogenous with respect to aspects of units, treatments, observations, and settings that are theoretically expected to be irrelevant to the treatment–outcome relationship. To the degree that the results of the experiment are consistent across different types of Units, different manipulations of the independent variable (Treatments), different measures of the dependent variable (Observations), and different types of Settings, the researcher can conclude that generalization of the findings is not limited.

DISCRIMINANT VALIDITY. Basic social psychological theory and theories of programs (Lipsey, 1993; West & Aiken, 1997) have specified the processes through which a treatment is expected to have an effect on the outcome. These theories identify specific Treatment constructs (causal agents) that are supposed to affect specific Observation constructs (dependent variables). For example, the specific Treatment of frustration, defined as the blockage of an ongoing goal-directed behavior, is hypothesized to lead to increases in aggression. Other similar treatments that do not involve goal blockage, such as completing an exciting task, should not produce aggression. Similarly, given the focus of the hypothesis on the construct of aggression, the researcher should be able to show that frustration does not lead to other emotion-related responses, such as depression or euphoria. To the extent that the causal agent of the Treatment (here, frustration) is shown to match the hypothesized construct and the class of Observations (here, aggression) affected by the Treatment matches those specified by the theory, claims for understanding the causal relationship are strengthened. This same approach can be taken to Units and Settings when hypotheses identify specific classes of units or settings over which the causal effect will generalize.

CAUSAL EXPLANATION. To the extent that we can support a causal explanation of our findings and rule out competing explanations, the likelihood of generalization is increased. The causal explanation distinguishes the active from the inert components of our treatment package and provides an understanding of the processes underlying our phenomenon of interest. These features permit us to specify which components need to be included in any new experimental context. This principle has long been at the heart of basic experimental work in social psychology with its focus on the articulation and ruling out of competing theoretical explanations. More recently, both basic and applied social psychologists have used mediational analysis (Baron &

Kenny, 1986; MacKinnon, 1994; West & Aiken, 1997) as a means of probing whether their favored theoretical explanation is consistent with the data. To the extent that the data support the favored theoretical explanation, it can provide strong guidance for the design, implementation, and evaluation of future programs. However, mediational analysis does not automatically rule out other competing explanations for the observed effects. To the extent that these alternative causal explanations (a) are plausible and (b) make different predictions when new units (participants), new treatments, new observations, or new settings are studied, generalization of the findings of an experiment will potentially be limited.

EMPIRICAL INTERPOLATION AND EXTRAPOLATION. For ease of presentation, this chapter has focused on the comparison of two treatment conditions, typically treatment versus control (or comparison) groups. Although experiments are conducted in social psychology with more than two levels of each separate treatment variable, such experiments are not common, particularly in applied social research. In general, a high dose of the treatment (e.g., a smoking prevention program carried out over 3 years; Evans et al., 1981) is compared with a no treatment (or minimal treatment) control group. This practice yields an estimate of the causal effect that is limited to the specific implementations of treatment and control groups used in the experiment.

Occasionally, parametric experiments or dose-response (response surface) experiments involving one or more dimensions (Box & Draper, 1987; Smith, this volume, Ch. 2; West, Aiken, & Todd, 1993) can be conducted. Such experiments help establish the functional form of the relationship between the strength of each treatment variable and the outcome of interest. Once the functional form of the dose-response curve (or response surface) is known, causal effects for the comparison of any pair of treatment conditions can be estimated through interpolation.

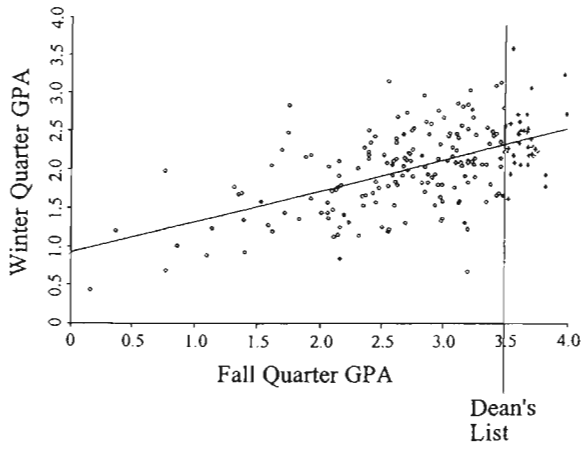
In the absence of such dose-response curves, caution must be exercised in generalizing the magnitude of causal effects very far beyond the specific levels of the treatment and control groups implemented in a particular experiment. The functional form of dose-response relationships may be nonlinear. Complications like threshold effects, the creation of new processes (e.g., psychological reactance if an influence attempt becomes too strong), and the influence of interactions with other background variables become increasingly likely as the gap between the treatments studied and those to which the researcher wishes to generalize increases. To the extent we are extrapolating beyond the range of units, treatments, observations, or settings used in previous research, our generalization of estimates of treatment effects become increasingly untrustworthy.

SUMMARY. Traditional social psychological perspectives on generalization (e.g., Aronson et al., 1998; Berkowitz & Donnerstein, 1982) have relied nearly exclusively on the single principle of causal explanation, making interpolation and extrapolation of causal effects difficult in the absence of a formal specification of the Units, Treatments, Observations, and Settings addressed by the theory. Cook's five principles add other criteria beyond causal explanation that help identify when generalization of the causal effects to the UTOS of interest is possible.[10] Cook (1993) and Shadish et al. (in press) also noted that these same five principles can also be applied to meta-analyses of entire research literatures.

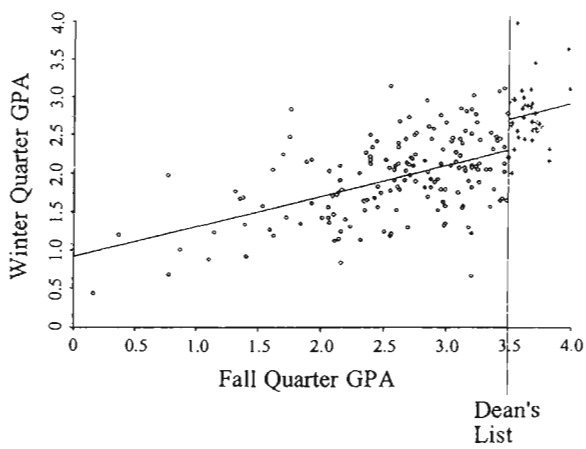
## QUASI-EXPERIMENTAL DESIGNS

The randomized experiment is nearly always the design of choice for testing causal hypotheses about the effects of a treatment. In many real-world contexts, however, randomization may be precluded by ethical, legal, practical (logistic), or policy concerns. In such cases, it is frequently possible to develop alternative quasi-experimental designs that share many of the strengths of the randomized experiment in reaching

[10] Some social psychologists concerned primarily with basic research argue that constancy of the direction of the causal effect is all that is needed for successful generalization. Taking a position based on traditional null hypothesis significance testing, they argue that whether the effect of a treatment is large, moderate, small, or even tiny in magnitude makes little difference so long as it is in the predicted direction (but see Greenwald, Pratkanis, Leippe, & Baumgardner, 1986; Reichardt & Gollob, 1997). On the other hand, some influential social psychological theorists (e.g., McGuire, 1983) have emphasized the contextual nature of social phenomena and much of the basic research conducted in social psychology during the past four decades has emphasized the finding of predicted interaction effects, ideally of a disordinal or crossover form. In contrast, in many areas of applied research, the magnitude of effects is currently viewed as more important. For example, an expensive job training program that led to a consistent mean increase of $100 per year in the participants' annual salary would be deemed to be ineffective, no matter how consistently the effect was obtained or how tiny the associated $p$-value of the statistical test (e.g., $p < .00001$).




(a)




(b)

valid estimates of the causal effect of a treatment (Cook & Campbell, 1979; Reichardt & Mark, 1997). Each of these designs involves manipulation of the independent variable by the experimenter or other entities, pretest and posttest measurement, and design or statistical controls that attempt to address plausible threats to internal validity.

### Regression Discontinuity Design

One of the strongest alternatives to the randomized experiment is the regression discontinuity design. The regression discontinuity design can be used when treatments are assigned on the basis of a quantitative measure, often a measure of need or merit. Following Reichardt and Mark (1997), we term this measure the *quantitative assignment variable*. For example, at some universities, entering freshmen who have a verbal scholastic aptitude test (SAT) score below a specified value (e.g., 380) are assigned to take a remedial English course, whereas freshman who have a score above this value are assigned to a regular English class (Aiken, West, Schwalm, Carroll, & Hsuing, 1998). The outcome of interest is the students' performance on a test of writing skill. Similarly, children who reach their sixth birthday by December 31 of the school year are assigned to begin first grade, whereas younger children do not begin school (Cahan & Davis, 1987). The outcome of interest is the children's level of performance on verbal and math tests taken the following Spring. Union members are laid off from work on the basis of their number of years of seniority. Mark and Mellor (1991) compared the degree of hindsight bias (retrospective judgments of the perceived likelihood of layoffs) among laid-off workers (< 20 years seniority) and workers who survived layoffs (20 or more years of seniority). The central feature of each of these examples is that participants are assigned to treatment or control conditions solely on the basis of whether they exceed or are below a cutpoint on the quantitative assignment variable and that an outcome hypothesized to be affected by the treatment is measured following treatment. This assignment rule meets the objections of those critics of randomized experiments who believe that potentially beneficial treatments should not be withheld from the neediest (or most deserving) participants.

To understand the strengths and weaknesses of this design, let us review a classic regression discontinuity study by Seaver and Quarton (1976) in some detail. Seaver and Quarton were interested in testing the hypothesis that social recognition for achievement leads to higher levels of subsequent achievement. A random sample of undergraduate students who completed the Fall and Winter quarters in a large university constituted the participants. As at many universities, those students who received a Fall quarter grade point average (GPA) of 3.5 or greater were given formal recognition by being placed on the Dean's list, whereas students who did not attain this GPA were not given any special recognition. The grades for all students were recorded at the end of the Winter quarter as a measure of subsequent achievement. Seaver and Quarton estimated that the awarding of Dean's list led to a 0.17-point increase in the students' GPA, the equivalent of a full grade higher in one 3 hour course during the next term.

To understand how the regression discontinuity design works, it is useful to consider several possible outcomes patterned generally after those of the Seaver and Quarton (1976) study. Each of the outcomes presented in Figure 3.3 is based on simulated data from 200 students, with about 17% of the students having Fall quarter GPAs of 3.5 or better (Dean's list).

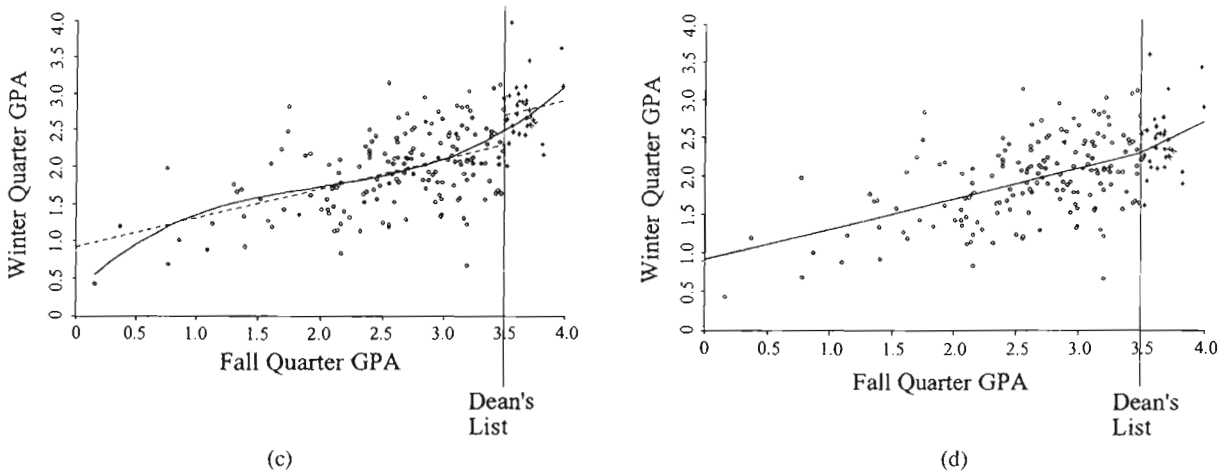



**Figure 3.3.** Illustration of possible outcomes in the regression discontinuity design.
(A) No effect of treatment.
(B) Positive effect of treatment.
(C) Curvilinear functional form.
(D) Nonparallel regression lines in treatment and control groups.

*Note.* In each panel, the relationship between students' Fall quarter and Winter quarter GPAs is indicated. The fit of a line represents the fit of a Fall GPA cutoff for Dean's List. o represents Control student; + represents a student on Dean's list. In Panel (A), representing no treatment effect of Dean's list, a single regression line fits the data for both the Dean's list and Control students. In Panel (B), the vertical distance between the regression lines for the Dean's list and for the Control students (the discontinuity) represents the treatment effect. In Panel (C), the solid line represents the fit of a curvilinear relationship between students' Fall and Winter quarter GPAs; the dotted line represents the treatment effect that would result if this relationship were specified as being linear. In Panel (D), the slope of the regression line for the Dean's list students is steeper than the slope of the regression line for the Control students (Treatment × Pretest interaction).

Figure 3.3 illustrates four possible outcomes of the study, of which outcomes (A) and (B) are important in the present context.

Outcome (A) illustrates a case in which the treatment, Dean's list, has no effect on Winter quarter GPA. Note that we have a single regression line that characterizes the full range of the data. The slope of the regression line indicates only that there is a strong positive relationship (correlation) between Fall quarter and Winter quarter GPA. In contrast, outcome (B) illustrates the general pattern of results found by Seaver and Quarton (1976). Here, the same regression line as in outcome (A) holds for students who have Fall GPAs below 3.5. However, for those students who have GPAs of 3.5 or above and are awarded Dean's list, the regression line is now elevated. At the cutpoint of 3.5 for Dean's list, we see that the difference in the levels (intercepts) of the two regression lines changes by 0.17, which represents Seaver and Quarton's estimate of the treatment effect. This discontinuity in the two regression lines can potentially be interpreted as a treatment effect.

Statistically, the treatment effect is estimated through the use of the following regression equation:

$$Y = b_0 + b_1X + b_2T + e. \quad (1)$$

In this equation, $Y$ is the outcome variable, here Winter quarter GPA; $X$ is the quantitative assignment variable, here Fall quarter GPA; and $T$ is a dummy code which has a value of 1 if the participant receives the treatment, here the social recognition of Dean's list, and a value of 0 if the participant does not receive treatment. Each of the $b$s is a regression coefficient that may be estimated by any standard statistical package (e.g., SPSS; SAS). The regression coefficients are most easily interpretable if the data are rescaled so that $X$ has a value of 0 at the cutpoint (i.e., $X = \text{GPA}_{\text{Fall}} - 3.5$). In this case, $b_0$ is the predicted value of Winter quarter GPA for participants at the cutpoint (3.5) who are in the control group (non-Dean's list), $b_1$ is the slope of the regression line (i.e., the predicted amount of increase in the Winter quarter GPA corresponding to a 1-point increase in the Fall quarter GPA), and $b_2$ represents the treatment effect. The test of $b_2$ informs us whether being on the Dean's list led to a significant increase in GPA the following quarter. Finally, $e$ is the residual (error in prediction). Readers wishing a more thorough presentation of the statistical procedures should see Aiken and West (1991, chapter 7); Cohen and Cohen (1983); Reichardt, Trochim, and Cappelleri (1995); or Trochim (1984).

The ideas behind the regression discontinuity design are straightforward, but the design is counterintuitive because it explicitly violates a usual canon of research: The treatment and control groups should be as equivalent as possible at pretest. Instead, this design takes advantage of the known rule for assignment to treatment (Dean's list: Fall GPA $\geq$ 3.5) and control groups (no recognition: Fall GPA $<$ 3.5). A statistical adjustment is performed that permits comparison of the levels in the two groups at the same value on the quantitative assignment variable[11] (GPA = 3.5). The estimate of the treatment effect is now conditioned on the participant's Fall GPA.

**STATISTICAL ASSUMPTIONS OF THE REGRESSION DISCONTINUITY DESIGN.** Because of its strong reliance on statistical adjustment to yield proper estimates of treatment effects, the regression discontinuity design requires two statistical assumptions in addition to those required for the randomized experiment. The ability to check these assumptions will be much greater in studies with large sample sizes, in which the cutpoint is not too extreme.

CORRECT SPECIFICATION OF FUNCTIONAL FORM. Researchers in social psychology rarely have strong a priori theoretical or empirical bases for specifying the form of the relationship between two variables. Consequently, we normally assume that the form of the relationship between the two variables is linear, as was done in the regression equation (Equation 1). To the degree that the form of the relationship between Fall quarter and Winter quarter GPA in our example is not well approximated by a straight line, the estimate of the treatment effect based on Equation 1 may be biased. For example, Cook and Campbell (1979, p. 140) presented data suggesting that a single curvilinear relationship without a treatment effect could fit the Seaver and Quarton (1976) data equally as well as the more typical linear regression equation (Equation 1). (Figure 3.3, Panel C illustrates this issue.) Consequently, it is important for researchers to examine the robustness of the results as the functional form of the relationship between the quantitative assignment and outcome variables is varied.

Two simple probes of the functional form may be performed. First, scatterplots of the relationship between the quantitative assignment variable and (a) the outcome measure and (b) the residuals should be carefully examined. As is usual in regression analysis, evidence of outliers, nonconstant variance of the residuals across the range of $X$, and nonnormality of the residuals suggest potential problems in the specification of the regression model (see R. D. Cook & Weisberg, 1994; McClelland, this volume, Ch. 15). Of particular importance in the context of the regression discontinuity design is the existence of systematic positive or negative residuals from the regression line near the cutpoint, suggesting a potentially major problem in the estimation of the treatment effect. This examination is greatly facilitated by the use of modern graphical techniques, such as the fitting of lowess curves (Cleveland, 1993) separately to the treatment and the control groups. Lowess curves are "nonparametric" curves that may be used to describe the functional form of the relationship in the sample. Second, the regression equation may be estimated using a nonlinear functional form. Following Trochim (1984) and Reichardt et al. (1995), higher order polynomial terms (e.g., $X^2$) are typically added to the regression equation and tested for significance. Alternatively, the data may be transformed to achieve linearity (Daniel & Wood, 1980), or other parametric or nonparametric regression equations suggested by the lowess curve may be estimated (Daniel & Wood, 1980; Hastie & Tibshirani, 1990). To the extent that similar estimates of the treatment effect are found, or the alternative specifications of the model fit the data less well than Equation (1), the possibility that a misspecified functional form accounts for the results can be minimized.

NO TREATMENT $\times$ PRETEST INTERACTION. Researchers normally make the assumption that the regression lines will be parallel in the treatment and control groups. However, this need not be the case, as is depicted in Figure 3.3, Panel D. In this example, there is no discontinuity in the regression line at the cutpoint of 3.5, but the slope of the regression line is slightly steeper above this value. This example illustrates one form of a Treatment $\times$ Pretest quantitative assignment variable interaction that may occur.

Treatment $\times$ Pretest interactions represent another variant of misspecification. Once again, they may be detected from careful examination of scatterplots and plots of residuals. If a Treatment $\times$ Pretest interaction

[11] In terms of RCM, the expected value of the treatment group is compared with the expected value of the control group, conditioned on the specific value of the quantitative assignment variable at the cutpoint (Rubin, 1977). If the quantitative assignment variable is the sole basis on which participants are given the experimental versus comparison treatments, then this difference provides an unbiased estimate of the treatment effect.

is suspected, a new term, $XT$, representing the interaction may be added to Equation (1), resulting in Equation (2):

$$Y = b_0 + b_1X + b_2T + b_3XT + e. \quad (2)$$

Once again, the $b_2$ coefficient estimates the treatment effect, and the $b_3$ coefficient provides information about the magnitude of the difference in the slopes in the treatment and control groups. Note that $X$ must be rescaled so that it has a value of 0 at the cutpoint if $b_2$ is to be easily interpretable (Aiken & West, 1991).

When significant changes in slope are detected, researchers need to be cautious in their interpretation of the results. If there is no discontinuity between the two regression lines (i.e., no treatment main effect), differences in slope are not usually interpretable. The alternative explanation that there is no treatment effect, but rather a nonlinear relationship between the quantitative assignment variable and the outcome variable cannot be easily ruled out. However, when there is a substantial discontinuity between the two regression lines at the cutpoint, such an alternative explanation becomes considerably less plausible, and the treatment effect can be directly interpreted at the cutpoint (Trochim, Cappelleri, & Reichardt, 1991). As with any interaction, the estimate of the treatment effect is conditional: It would be different at any other potential cutpoint. Statistically, treatment effects at other cutpoints are easily estimated by rescaling the value of $X$ to be 0 at the new cutpoint and examining the new $b_2$ effect; however, extrapolation of estimates of treatment effects to other cutpoints makes two strong assumptions (Cochran, 1957). First, the regression model is a close approximation of the "true" regression model in the population. Second, the same true regression model holds for all values of the quantitative assignment variable. Thus, extrapolation yields treatment effect estimates that are less precise and less credible than the estimates at the actual cutpoint on the quantitative assignment variable.

**SELECTION ISSUES IN THE REGRESSION DISCONTINUITY DESIGN.** The regression discontinuity design assumes that participants are assigned to treatment and control groups solely on the basis of the cutoff score on the quantitative assignment variable. Any other influences that affect assignment represent a potentially serious misspecification of the model. Paralleling our earlier review of the randomized experiment, breakdowns in treatment assignment give rise to the possibility that there may be important differences in the people participating in the treatment and control groups after statistical adjustment for the quantitative assignment variable, even in the absence of treatment. We term such influences participant selection issues; they can arise in the regression discontinuity design in at least 3 ways.

1. If the population to which a desirable (or undesirable) treatment is being offered is aware of the cutpoint, participants may decide to enroll in the study based in part on their perception of the likelihood that they will receive the treatment. Meier (1985) reported that the Salk polio vaccine was tested in some states using a simple version of the regression discontinuity design in which the child's age was the assignment variable. However, many parents knew that children in a specified age range would receive the experimental polio vaccine. Gilbert, Light, and Mosteller (1975) documented how refusals of parents of children in this age range to permit their children to participate in the study led to a large bias in the estimates of the effectiveness of the vaccine.
2. Practitioners may "adjust" the scores of individuals who are just below (or above) the cutpoints so that these individuals may receive a desired treatment, a problem Campbell (1969) termed "fuzzy assignment rules." Teachers may give their favorite students better grades so these students receive academic honors; welfare workers may understate a family's income so that a child may receive special medical or educational treatment. Those in charge of treatment allocation may also take into account other unquantified factors, such as letters of recommendation or interviews with the candidates.
3. Following the pretest, participants may drop out of the treatment and control groups. In such cases, participant characteristics that may not have been measured at pretest, such as their interest in the topic area of the study, may determine in part whether they complete the study. As in the randomized experiment, the estimate of the treatment effect will be biased to the extent that attrition is substantial, the characteristics of attriters and completers differ in the treatment and control conditions, and these characteristics are related to the measured outcome variable.

Many of the remedies for selection issues in the regression discontinuity design are extensions of those used in the randomized experiment. In terms of design approaches, not announcing the cutpoint until after participants have been assigned to conditions masks

(blinds) the assignment rule from both participants and personnel in the institutional setting and thereby helps minimize the first two selection issues. Trochim and Cappelleri (1992) have suggested that the assignment rule used in the regression discontinuity design addresses the third selection issue because it is more sensible to participants; however, little evidence exists with which to evaluate their claim of reduced attrition. A variety of simple (e.g., dropping participants within a narrow range around the cutpoint; Mohr, 1988) and complex econometric approaches have attempted to provide more adequate adjustments when fuzzy assignment appears to have taken place (see Trochim, 1984, for a review). Statistical work on issues related to other selection issues is less well-developed. However, many of the statistical techniques suggested for the problems of treatment noncompliance and attrition in the randomized experiment can potentially be extended to the regression discontinuity design. These approaches make the strong assumptions that the quantitative assignment variable provides the only basis for treatment assignment and that the functional form of the relationship between the quantitative assignment variable and the outcome variable has been correctly specified.

#### OTHER ISSUES

STATISTICAL POWER. The regression discontinuity design will typically have considerably lower power than the randomized experiment, with the degree to which power is reduced being dependent on the extremity of the cutpoints and the magnitude of the correlation between the quantitative assignment variable and the posttest (Cappelleri, 1990; Cappelleri, Darlington, & Trochim, 1994). Goldberger (1972) estimated that, in properly specified models in which the quantitative assignment variable and the posttest had bivariate normal distributions and a relatively strong correlation, 2.75 times as many participants would be required using the regression discontinuity design to achieve the same level of statistical power as in the randomized experiment. The lesson is clear: Researchers planning regression discontinuity designs will need to use relatively large sample sizes, both to provide a test of the treatment that has adequate statistical power and to probe the statistical assumptions underlying the test.

CAUSAL GENERALIZATION. Drawing on Cook's (1993) five principles of causal generalization discussed earlier, we see that generalization of results using the regression discontinuity design is limited relative to the randomized experiment by the use of one cutpoint. Ideally, generalization of the treatment effects should be restricted to values that are close to the original cutpoint. Often this will be sufficient: Regression discontinuity designs are typically conducted with the populations and in the settings of interest and are thus high in external validity. The cutpoints used are also often those for which there are supporting data (e.g., cutoffs for clinical levels of high blood pressure) or strong historical tradition for their use (e.g., Dean's List = 3.5 GPA). Thus, this limitation is often an issue more in theory than in practice because of the restricted range of generalization that is sought.

**SUMMARY.** The regression discontinuity design provides an excellent approach when participants are assigned to conditions on the basis of a quantitative measure. As Marcantonio and Cook (1994) noted, the regression discontinuity design is one of two quasi-experimental designs that stand out "because of the high quality of causal inference they often engender" (p. 134). When its assumptions are met, it rules out most of the threats to internal validity and has good external validity for much applied work. On the negative side, it is considerably lower in power than the randomized experiment. The central concerns of the design focused around a number of selection-related issues and the correct specification of the functional form. More complex variants of the regression discontinuity design can be implemented: Multiple pretest measures may be used to assign participants to treatment, multiple treatments may be delivered with multiple cutpoints, multiple pretest measures may be used as additional covariates, and multiple outcome variables may be collected (see Judd & Kenny, 1981; Shadish, et al., in press; Trochim, 1984). In addition, the regression discontinuity design can be supplemented with a randomized tie-breaking experiment around the cutpoint to provide even stronger inferences (Rubin, 1977; Shadish, et al., in press; see Aiken et al., 1998, for an empirical illustration of combining these designs).

### Interrupted Time Series Design

The interrupted time series design is the second of the two quasi-experimental designs that Marcantonio and Cook (1994) highlighted as engendering a high quality of causal inference. In the basic interrupted time series design, measurements of the outcome variable are collected at equally spaced intervals (e.g., daily; yearly) over a long period of time. An intervention is implemented at a specific point in time; the

intervention is expected to affect the outcome variable measured in the series. Time series analysis permits the researcher to identify changes in the level and slope of the series that occur as a result of the intervention.[12] In terms of the RCM, the expected value of the treatment group is compared with the expected value of the control group, conditioned on the specific point in time at which the treatment is introduced. If (a) time is assumed to be a good proxy for the actual rule upon which treatment and control conditions are assigned (Judd & Kenny, 1981) and (b) the correct functional form of the relationship between time and the outcome variable is specified, then controlling for time will lead to an unbiased estimate of the treatment effect. If time is not an adequate proxy, then estimates of the magnitude of the treatment effect may be biased.

Interrupted time series designs have typically been utilized in two different research settings. First, time series designs provide an excellent method of conducting naturalistic studies of the effects of a change in law or a social innovation. To cite two examples, West, Hepworth, McCall, and Reich (1989) used this design to test whether the introduction of a law requiring a 24-hour mandatory jail term for driving under the influence of alcohol led to a subsequent decrease in fatal traffic accidents. Hennigan et al. (1982) used this design to test the hypothesis that the introduction of television broadcasting in U.S. cities in the late 1940s and early 1950s would lead to subsequent increases in crime rates, particularly violent crime rates. Second, time series designs may be used to test the effects of interventions in single-subject designs. This approach has primarily been followed by experimental psychologists in the Skinnerian tradition and clinical researchers who wish to make causal inferences at the level of the individual participant (Crosbie, 1993; Franklin, Allison, & Gorman, 1996; Horne, Yang, & Ware, 1982; Sidman, 1960). For example, Gregson (1987) used time series methods to understand individual differences in the pattern of response to treatment in 9 participants with chronic headaches. These participants reported the intensity and duration of their headaches on a daily basis over an extended period of time. In each of these examples, the treatment in this design was assigned a priori on the basis of time (or after a fixed number of observations[13]), and possible treatment effects may be inferred from a change in the level of the behavior that has occurred between the baseline and treatment periods.

Figure 3.4 illustrates some of the different types of effects that can be detected using interrupted time series designs. In Panel A, there is no effect of the intervention. In Panel B, the intervention leads to a dramatic decrease in the level of occurrence of the behavior. In Panel C, the intervention leads to an immediate effect of the intervention, followed by a slow decay of the treatment effect over time. Finally, in Panel D, there is a permanent change in both the level and the slope of the series following the intervention. Each of the results depicted in Panels B–D are potentially interpretable as treatment effects, given appropriate statistical analyses and attention to potential threats to internal validity.

From consideration of Figures 3.3 and 3.4, readers may have detected a strong similarity between the regression discontinuity design and the interrupted time series design. In both designs, treatment is assigned beginning at some specific point represented on the X-axis. Any alternative explanations of the results need to provide an account of why the level, slope, or both of the series change at that specific point. Note, however, that there is one major conceptual difference between the two designs: In the interrupted time series design, participants are assigned to the treatment or control group on the basis of time; in the regression discontinuity design, participants are assigned to the treatment or control group on the basis of their score on the quantitative assignment variable. As we show below, this change from pretest scores to time on the X-axis greatly complicates the statistical analysis and leads to a different set of potential threats to internal validity.

**AN OVERVIEW OF STATISTICAL ANALYSES.** To illustrate time series analysis, consider a classic study by McSweeny (1978). McSweeny was interested in the effects of the introduction of a monetary charge for calls to directory assistance. He obtained the monthly records of the number of calls made to Cincinnati Bell directory assistance for the period 1962 to 1976. Beginning in March 1974, and announced with considerable publicity, Cincinnati Bell began charging 20 cents

[12] Although less often hypothesized by social psychologists, changes in the variance of the series or in cyclical patterns in the series following an intervention can also be detected. Larsen (1989) has discussed some ways in which cyclical patterns of variables such as mood and activity levels may be important in human social behavior.

[13] Franklin et al. (1996) discussed some of the inferential problems that occur when other assignment rules (e.g., the participant's level of responding reaches asymptote) are used to determine when treatments are introduced or withdrawn.

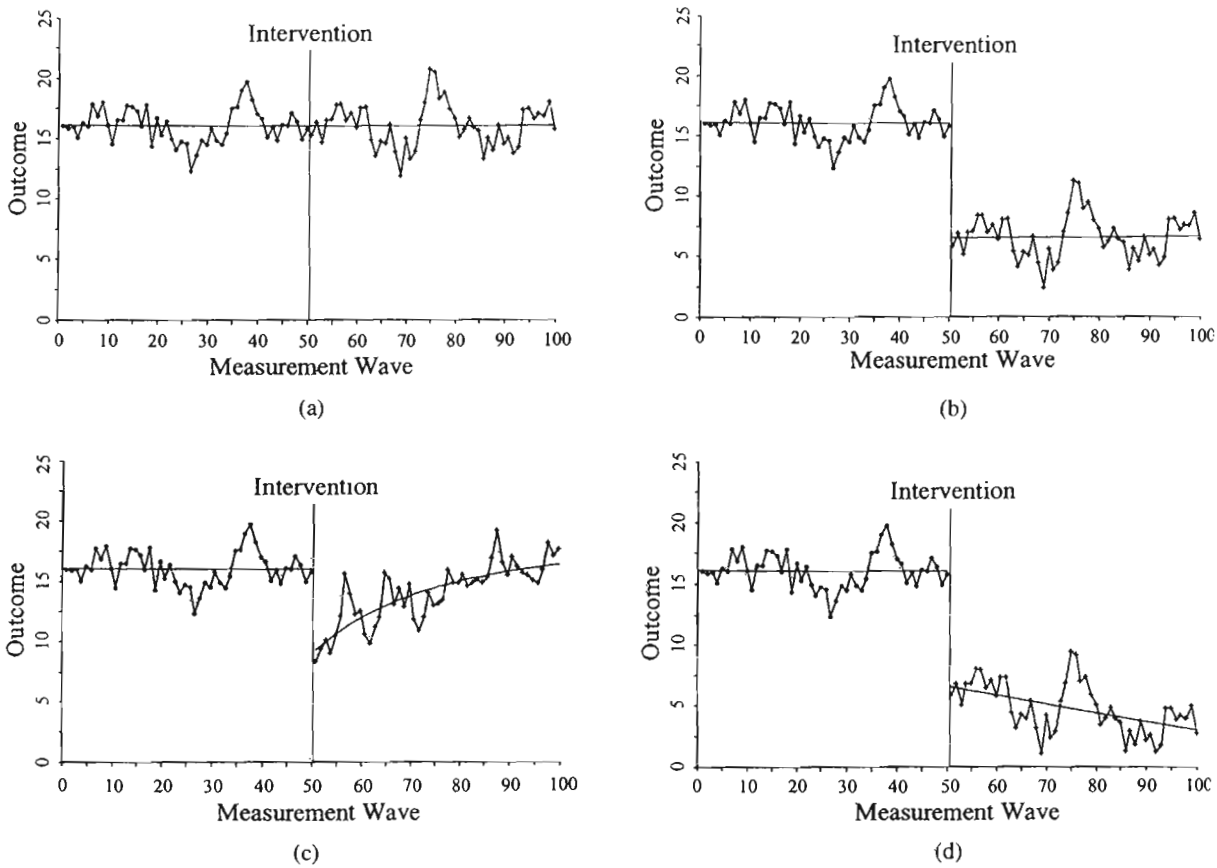



**Figure 3.4.** Illustration of possible outcomes in the interrupted time series design (A) No effect of intervention. (B) Intervention decreases the level of the outcome variable. (C) Intervention produces an immediate drop in level followed by a slow return to the baseline level of the series. (D) Intervention produces both an immediate drop in the level and a change in the slope of the series.

*Note.* The same outcome is assessed at each wave of measurement of the study. The vertical line indicates the point at which intervention is introduced (between measurements 50 and 51).

for each directory assistance call. As can be seen in Figure 3.5, the number of calls made to this service generally increased up to the point of the intervention (20-cent charge). At this point, the number of calls made to this service showed a large drop, then began increasing again at about the same rate as before.

The basic regression equation used in time series analysis closely parallels Equation 1, the equation used to estimate the treatment effect in the basic regression discontinuity design. To test the effect of the introduction of the charge for directory assistance, the following regression equation would be used:

$$Y = b_0 + b_1 T + b_2 Z + e. \quad (3)$$

Here, $Y$ is the number of calls made to directory assistance each month (outcome variable), $T$ is the month in the series, and $Z$ is the intervention effect, again dummy coded with 1 representing treatment (20-cent charge) and 0 no treatment control (no charge). Once again, the $b$s are regression coefficients, with $b_0$ representing the intercept (the predicted value of $Y$ when $T$ and $Z = 0$) and $b_1$ the slope of the regression line, and with $b_2$ representing the treatment effect (the change in the level of the series at the point of the intervention). For purposes of interpretability, $T$ is typically rescaled so that it has a value of 0 at the point of implementation of the treatment.[14]

Although this equation appears to be nearly identical to Equation 1, the use of time rather than the

[14] Using this rescaling, $b_0$ always gives the predicted value for the control series at the point of the intervention. In more complex models, such rescaling can be particularly important.

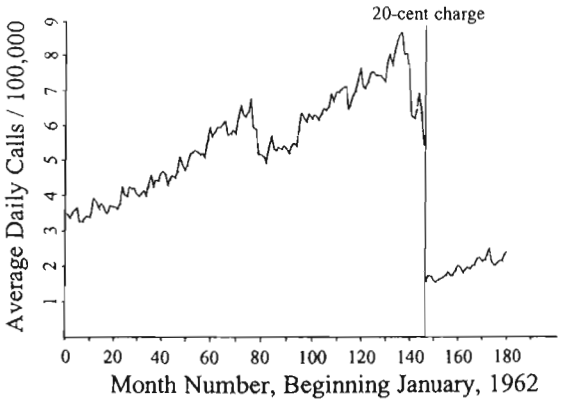



**Figure 3.5.** Intervention effect: Introduction of a charge for directory assistance.

*Note.* The series represents the average daily number of calls to directory assistance per 100,000 total calls. The vertical line indicates the point (the beginning of month 147 of the series) when the charge for directory assistance was introduced. The effect of this charge is shown by the large vertical drop in the level of the series at the point of intervention.

pretest score introduces major complications into the statistical analysis. In the analysis of time series data, three statistical problems must be adequately addressed or the regression equation will be misspecified. First, any long-term linear or curvilinear trends over time must be properly represented. In the present example, the long-term linear increase over time in the series is represented by the $b_1T$ term in Equation 3. Second, time series data, particularly those in which the data are collected on a hourly, daily, or monthly basis, frequently contain cycles. For example, people's moods often change according to a regular weekly cycle, with particularly large differences being found between the weekend and weekdays. Again, these cycles must be detected, and terms representing the cycles must be added to the regression equation to avoid misspecification (see West & Hepworth, 1991). Finally, when data are collected over time, adjacent observations are often more similar than observations that are further removed in time. For example, the prediction of today's weather from the previous day's weather is, on average, far better than the prediction of today's weather from the weather 7 days ago. This problem, known as serial dependency, implies that the residuals (the *e*s) in Equation 3 will not be independent (Judd & Kenny, 1981; West & Hepworth, 1991). Failure to properly specify the long-term trends, the cycles, or both in the model may lead to model misspecifications and consequently to bias in the estimate of the treatment effect. Failure to adequately represent the serial dependency leads to incorrect standard errors and consequently to incorrect significance tests and confidence intervals (Box, Jenkins, & Reinsel, 1994; Chatfield, 1996; Judd & Kenny, 1981).

Time series analysis offers a variety of graphical displays and statistical tests that are useful in the detection of trends, cycles, and serial dependency. Time series techniques are most accurate in large samples where the number of time points is 100 or more, although even in large samples there can be problems in correctly identifying the pattern of serial dependency (Velicer & Harrop, 1983). In smaller samples, there is a high likelihood that several models will adequately fit the data, and the possibility exists that each of these models may be associated with a substantially different estimate of the treatment effect (see Velicer & Colby, 1997). This result occurs because time series analysis requires a series of preliminary tests to determine whether the model is correctly specified; these preliminary tests require large numbers of observations to distinguish sharply between different possible models. Among the techniques that may be used to identify potential problems with time series models are graphical techniques for visualizing trends and cycles (see Cleveland, 1993) and plots of specialized statistics (e.g., spectral density plots; autocorrelograms) to detect cycles and serial dependency. McCleary and Hay (1980), Judd and Kenny (1981, chapter 7), West and Hepworth (1991), and Velicer and Colby (1997) presented good introductions to the ideas of time series analysis. Chatfield (1996), Box, Jenkins, and Reinsel (1994), and Velicer and MacDonald (1984) presented more in-depth treatments of the traditional and newer statistical procedures for the identification and testing of time series models.

**THREATS TO INTERNAL VALIDITY.** As in the regression discontinuity design, for threats to internal validity to be plausible, they must offer an explanation of why a change in the level, slope, or both of the series occurred at the point of the intervention (Shadish et al., in press). Three generic classes of explanations that are potential threats to the internal validity of the design can be identified; however, the plausibility of each of these threats will depend on the specific research context.

As an illustration, consider the example of a new smoking prevention program implemented at the beginning of a school year. Based on school records, the total number of students who are cited by school personnel for smoking each month on school grounds for 5 years prior to and 5 years after the beginning of the intervention provide the data for the time series.

HISTORY. Some other event that could also be expected to decrease smoking may occur at about the same time that the smoking prevention program is initially implemented. For example, the community may remove cigarette machines or institute fines for selling cigarettes to minors, making access to cigarettes more difficult.

SELECTION. The population of the students in the school may change in the direction of having a greater proportion of nonsmokers. Parents who support the prevention efforts may hear about the planned intervention and make efforts to enroll their children in the school, whereas parents who support smokers' rights may transfer their children to other districts.

INSTRUMENTATION. Aspects of the record-keeping procedures may change, leading to decreased reports of smoking. Students caught smoking may no longer be written up for their first offense or decreased staff may be available for monitoring of the school's grounds.

Once again, these threats are only plausible if they occur at about the point of the intervention. For example, if the community were to remove cigarette machines 3 years after the implementation of the smoking prevention program, this action would not provide an alternative explanation of a drop in the number of students who are cited for smoking at the point of implementation of the program. Many of these threats are less plausible in time series designs using single participants, in which the experimenter has considerable control over the procedures than in less controlled, naturalistic designs testing the effects of new laws, programs, or new innovations on a population of participants.

**DESIGN ENHANCEMENTS.** The three potential threats to internal validity can be made less plausible through the addition of one or more design enhancements to the basic time series design. We consider three design enhancements that can often be used both in naturalistic tests of policy and single subject designs.

NO TREATMENT CONTROL SERIES. Comparable data may be available from another similar unit that did not receive treatment during the same time period. Recall that West et al. (1989) studied the effect of a new state law in Arizona that mandated a 24-hour jail term for drivers convicted of driving while intoxicated. They showed that this law led to an immediate 50% decrease in traffic fatalities in the city of Phoenix; however, the magnitude of this decrease declined over time following the intervention. In an attempt to rule out possible historical effects (e.g., unusually good weather; change in speed limit) coinciding with the implementation of the new law, West et al. also analyzed the traffic fatalities from a city (El Paso) in a nearby state that did not introduce a similar drunk driving law during the period of the study. The results showed no comparable change in traffic fatalities in El Paso at the same point in time (July 1982) when the law went into effect in Arizona. Ideally, the control series should be as similar as possible to the treatment series and similar data should be available for the same time period.

OTHER CONTROL SERIES. Sometimes data are available from another series that would not be expected to be influenced by the treatment, but which would be expected to be impacted by many of the same nontreatment influences as the treatment series of interest. Reichardt and Mark (1997) illustrated examples of such control series using data collected on the same participants in different settings or using different measures that are not expected to be affected by the treatment. Often such a series can be constructed by disaggregating a series into units in which little, if any, impact is expected versus units in which a large impact of the intervention is expected. In a classic study, Ross, Campbell, and Glass (1970; see also Glass, 1988) studied the introduction of a new intervention to decrease drunk driving in Great Britain. British police were equipped with breathalysers and were empowered to suspend the drivers license of anyone who was driving while intoxicated. Ross et al. divided the data into weekend evenings when the amount of alcohol use is normally high and weekday mornings when alcohol use is normally low. Consistent with their predictions, the implementation of the breathalyser program led to a large decrease in fatalities in the weekend evening hours, followed by a slow return to baseline. However, the implementation of the program did not lead to a detectable decrease during the weekday morning hours. The comparison of the two series helps rule out effects due to history and changes in instrumentation that should affect both series equally. Selection is an implausible threat in this case because it is unlikely that a country would experience appreciable in or out migration or changes in the population of drivers that were associated with a traffic safety campaign against drunk driving.

SWITCHING REPLICATIONS. In some cases, the same intervention may be implemented in more than one locale at different times. Time series analyses of the data from the different locales would be expected to show similar effects occurring at the point of intervention in each locale. For example, West et al. (1989) found that California implemented a law mandating a 24-hour jail term for driving under the influence of alcohol 6 months before the similar law was implemented in Arizona. Analysis of highway fatality data from San Diego showed a similar effect to that which occurred 6 months later in Phoenix – an immediate 50% reduction in highway fatalities followed by a slow decrease in the magnitude of the effect over time. The replication of the effect at a different point in time helps rule out other possible explanations associated with history (e.g., weather) and changes in instrumentation.

COMBINING DESIGN ENHANCEMENTS. As should be evident, multiple design features can be combined in a single study to strengthen further the causal inferences that may be made. For example, West et al. (1989) combined all three of the previously discussed design features in their studies of the effects of drunk driving laws. Ross et al. (1970) added a number of other control series to their basic time series design, together with a number of other design and measurement features that were designed to address specific threats to internal validity. When the researcher has control over treatment delivery, the design can be further strengthened by introducing and removing treatments following an a priori schedule. In well-designed time series designs, inferences can sometimes be made with a certainty approaching that of a randomized experiment.

**DELAYED EFFECTS.** Time series designs provide a relatively strong basis for causal inferences when the intervention produces immediate effects, but the strength of the causal inference may be weakened considerably when treatment effects are delayed – unless the length of delay in the effect is predicted ahead of time. As an illustration, consider again our example of the school-based smoking prevention program. Suppose that a gradual decrease in the number of students cited for smoking on school grounds began 1 year following the implementation of the program. Under such circumstances, it would be difficult to attribute these changes in the level and slope of the series to the prevention program.

True delayed effects normally occur in one of two ways. First, new policies often do not go abruptly into effect on the specific starting date of the intervention. New programs often require time for personnel to be trained before they fully go into effect; new innovations often require time to diffuse through society. Inferences from time series may be strengthened by collecting supplementary data that assess the extent of implementation of the intervention over time. Using such data to model a gradual implementation process can strengthen causal inferences and increase the power of the statistical tests in time series analyses. Hennigan et al. (1982) used data on the proportion of households with television sets following the beginning of television broadcasts to strengthen their potential causal inferences about the effects of television on crime rates. In addition, they observed that the pattern of increase in burglaries was nearly identical following the beginning of broadcasting in some American cities in 1948 and others in 1952 in their switching replication design.

Second, some interventions affect processes whose outcomes will only be evident months or years later. To cite two examples, a birth control intervention would be expected to affect birth rates about 9 months later. A nationwide school-based smoking prevention program would be expected to show effects on lung cancer rates beginning about 35 years after the intervention as the participants matured into the age range when lung cancer first begins to become manifest. In both cases theory provides a strong basis for expecting the effect to occur at a specified time lag or over a specified distribution of time lags following the intervention. In the absence of strong theory, similar strong causal inferences cannot be made. As previously noted, a decrease in school citations for smoking 1 year following the introduction of a smoking prevention program cannot be interpreted confidently as a program effect.

**STATISTICAL POWER.** Two issues of statistical power commonly arise in time series analysis. First, researchers often evaluate the effects of an intervention shortly after an intervention has been implemented, giving rise to a series with many preintervention, but few postintervention observations. Given a series with a fixed number of time points, the statistical power of the analysis to detect treatment effects is greatest when the intervention occurs at the middle rather than one of the ends of the series. Second, the existence of serial dependency has complex effects on calculations of statistical power. Within the preintervention series and within the postintervention series, to the degree that observations are not independent, each observation in effect counts less than 1, thus increasing the total number of observations needed to achieve a

specified level of statistical power. On the other hand, the use of the same participant or same population of participants throughout the series reduces the variability of the series relative to one in which different participants are sampled at each time point. The statistical power of a given time series analysis is determined in large part by the trade-off between these two effects in the specific research context.

**CONCLUSION.** Interrupted time series designs provide a strong quasi-experimental approach when interventions are introduced at a specific point in time so that time may be used as a proxy for the true model of treatment assignment. Such designs provide a strong basis for the evaluation of the effects of changes in social policy or new innovations that are implemented at a specific point in time. The basic time series design often allows researchers to credibly rule out several threats to internal validity: Any alternative explanation, to be plausible, must account for why the change in the series occurred at a particular point in time. Design enhancements, including the use of various control series, switching replications, and the introduction and removal of treatments, can further strengthen the causal inferences that may be made (see also Barlow & Hersen, 1984; Kratochwill & Levin, 1992). Given their ability to rule out alternative explanations through both design features and statistical adjustment, interrupted time series designs represent one of the strongest alternatives to the randomized experiment.

### Nonequivalent Control Group Design

The most commonly used quasi-experimental alternative to the randomized experiment is the nonequivalent control group design. In this design, a group of participants is given a treatment, or a "natural event" occurs to the group. A second (comparison) group that does not receive the treatment is also identified. Both groups of participants are measured both before and after the treatment. Randomization is not used to determine the treatment group assignment; rather the process through which participants end up in the treatment and comparison groups is not fully known and is therefore difficult to model statistically. To cite three examples, Lehman, Lampert, and Nisbett (1988) compared the level of statistical reasoning of advanced graduate students in psychology, which emphasizes training in statistics, with that of advanced graduate students in other disciplines (e.g., chemistry) that have substantially lower training emphasis in statistics. Coleman, Hoffer, and Kilgore (1982) and Murname, Newstead, and Olson (1985) compared the academic achievements of students who attended private schools with those who attended public schools. And, Martin, Annan, and Forst (1993) compared the subsequent arrest rates of individuals who were convicted of drunk driving and who were sentenced by one of two judges. One judge sentenced virtually all persons to a 2-day jail term, whereas the other (comparison) judge sentenced virtually all persons to non-jail alternatives (e.g., monetary fines, community service). In each example, measures of the important outcome-related variables of interest were collected in the treatment and control groups both prior to and following the intervention. The outcomes in the treatment and control group at posttest were compared after attempting to remove any differences between the groups that were observed at pretest.

The nonequivalent control group design appears to provide a straightforward way of investigating the causal effect of a treatment when randomization is not possible. However, direct (unadjusted) comparison of the treatment and control group means is only legitimate when a strong assumption is met: The treatment and control groups must be equivalent in terms of all important background characteristics at pretest. This assumption requires that both the mean pretest levels of the treatment and control groups are the same and that the rate of maturation (growth or decline) in the two groups is the same in the absence of treatment. Given random assignment to treatment and control groups, this assumption is normally met because the treatment group assignment is expected, on average, to be independent of all participant background characteristics. However, in the nonequivalent control group design, the treatment and the control groups must be presumed to be nonequivalent at pretest; researchers must also presume that measured (or unmeasured) participant characteristics at pretest will be related to treatment assignment. Following Rosenbaum (1995), we refer to variables assessed at pretest (or other background variables measured prior to treatment) as *covariates* and variables not assessed at pretest as *hidden variables*. Ruling out the possibility that the two groups may differ prior to treatment in terms of either measured covariates or hidden variables is the central task facing the researcher using this design. Major disputes have occurred in the literature over the "true effectiveness" of treatments (e.g., the Head Start program) that have been evaluated using this design. Researchers may legitimately retain considerable uncertainty as to whether preexisting differences between the treatment

and control groups on covariates and hidden variables have been adequately ruled out as potential explanations for observed differences on the outcome variable (e.g., see Barnow, 1973; Bentler & Woodward, 1978; Cicarelli, Cooper, & Granger, 1969; Magidson, 1977 for diverse analyses and reanalyses of the data from the original Head Start study). Because of this problem, researchers are well advised to add other design features (presented later in this section) to the basic nonequivalent control group design in order to reach more convincing causal inferences.

**STRATEGIES FOR EQUATING GROUPS.** From the standpoint of the RCM, the goal of the researcher using the nonequivalent control group design parallels that of the regression discontinuity design: The researcher must adequately model the mechanism by which participants were assigned to the treatment and control groups. When this goal is accomplished, the treatment effect estimate will be an unbiased estimate of the treatment effect in the population (Rosenbaum, 1984; Rosenbaum & Rubin, 1984). As in the regression discontinuity design, the estimate of the treatment effect is conditioned on the participant background variables that are related to treatment assignment.

As noted previously, the central problem in the nonequivalent control group design is that units (participants) are not assigned to the treatment and control groups according to any known rule. Researchers must use whatever information they have available to identify pretest or other background variables (e.g., socioeconomic status, prior educational or medical history) with which to model the process of selection into the treatment and control groups. The critical background variables are those that are related to both (a) selection into the two treatment groups and (b) the outcome variable in the population. Pretest variables that are only related to either (a) or (b) but not both do not bias treatment effect estimates (Berk, 1988).

Researchers use these pretest and background variables in an attempt to adjust statistically for differences between the two groups that existed prior to treatment. Several assumptions must be made in order for these statistical adjustment techniques to produce unbiased estimates of the causal effect. These include

1. All important covariates have been identified; there are no hidden variables.
2. Each of the covariates has been measured with perfect reliability, or statistical procedures (e.g., structural equation modeling) have properly adjusted for unreliability in the covariates.
3. The functional form (e.g., linear) of the relationship between the pretest variables and the outcome variable has been correctly specified.
4. The pretest maturation rates do not differ in the treatment and control groups or proper adjustments for differences in the pretest maturation rates have been successfully undertaken. In the basic nonequivalent control group design, such differences in maturation rates can be indicated by Group × Pretest interactions or unequal variances in the treatment and control groups at pretest or posttest.

When each of these assumptions is met, the nonequivalent control group design has in effect been made equivalent to the regression discontinuity design. The estimate of the treatment effect conditioned on the statistical model that properly adjusts for selection into the treatment and control groups will lead to an unbiased estimate of the causal effect in the population. All other covariates and hidden variables will then be unrelated to treatment condition, a condition that has been termed *strong ignorability* (Holland, 1986; Rosenbaum, 1984; Rosenbaum &Rubin, 1984). In practice, however, meeting these assumptions and thereby achieving strong ignorability is both extraordinarily difficult and fraught with uncertainty. Consequently, researchers cannot be confident in practice that their estimates of the treatment effect using the nonequivalent control group design are unbiased.

**SELECTING COVARIATES.** Statisticians and methodologists have considered a variety of possible methods of identifying important covariates that should be included in statistical models that attempt to adjust for pretest differences and achieve strong ignorability. Reichardt, Minton, and Schellenger (1986) suggested considering both (a) those variables that are theoretically expected to be related to selection into the treatment and control groups and (b) those variables that are known from the literature to be related to the outcome variables. Cochran (1965) and Rosenbaum (1995) have both suggested conducting exploratory analyses of all variables measured at pretest. Based on simulation work, Cochran suggested retaining any covariate for further consideration for which the $t$-value for the difference between the treatment and control groups was greater than 1.5 at pretest (see also Canner, 1984, 1991). This technique can be supplemented by a reexamination of the excluded covariates once the selection model has been developed to detect any remaining covariates that are

not adequately equated in the treatment and control groups (Rosenbaum, 1995).

Note that the pretest measure of the outcome variable of interest has a special status in the nonequivalent control group design. Reichardt (1979) and Rosenbaum (1995) have both noted that the best method of reducing sensitivity to unmeasured variables is to identify covariates that are (a) unaffected (i.e., not caused) by the treatment, but are (b) strongly related to the outcome variable. The pretest score on the outcome of interest, particularly when there is a short time lag between pretest and posttest, normally meets these requirements.

**STATISTICAL ADJUSTMENT TECHNIQUES.** A variety of statistical techniques can be used to analyze the data from the nonequivalent control group design. The techniques do not always produce the same result; they use different statistical procedures and consequently have different strengths and weaknesses. For example, only the second technique considered below, analysis of covariance with correction for unreliability, addresses the issue of measurement error. Nonetheless, each statistical technique has the same goal: To attempt to achieve strong ignorability by providing a statistical adjustment that properly removes prestest differences between the treatment and control groups.

ANALYSIS OF COVARIANCE. Analysis of covariance adjusts the posttest score for the *measured* pretest scores. Typically, adjustment is made only for the linear effect of the pretest scores on the posttest scores. No adjustment is made for variables that are not included in the analysis nor is the proper adjustment made for variables that are measured with error. When one covariate is used, unreliability will result in too little adjustment for the initial differences between treatment groups. When more than one covariate is used and one or more of the variables is less than perfectly measured, the outcome is less clear: Typically, unreliability results in too little adjustment, but proper adjustment, or even overadjustment, may occur in some cases. In a specific study, the estimate of the treatment effect may be too large, too small, or even just right.

ANALYSIS OF COVARIANCE WITH CORRECTION FOR UNRELIABILITY. In this procedure, an attempt is made to adjust the treatment effect based on an estimate of what the pretest scores would have been if they had been measured without error (Huitema, 1980). This adjustment is now typically performed using structural equation modeling programs such as EQS (Bentler, 1995) or LISREL (Jöreskog & Sörbom, 1993). In one variant of this procedure, each measured covariate is specified to be the result of an unmeasured true score and an error of measurement. The variance of the error of measurement is normally set equal to the product of the variance of the variable $\times$ (1 $-$ reliability of the measure), i.e., $\sigma^2[1 - \rho_{XX}]$; see Bollen, 1989. The test–retest correlation for a relatively brief time interval is normally preferred as the estimate of $\rho_{XX}$, although other measures of reliability may be preferred in some situations[15] (Campbell & Boruch, 1975; Campbell & Erlbacher, 1970; Linn & Werts, 1973). The outcome variable is then regressed on the true scores of each pretest variable and a dummy coded variable representing the treatment. The estimate then represents the effect of the treatment adjusted for the differences among the covariates at pretest, corrected for the measure of unreliability. A second important method of correcting for unreliability uses multiple measures of each pretest construct rather than a single measure that is adjusted with an estimate of unreliability. Using structural equation modeling, the multiple measures are used to provide a theoretically error-free estimate of each latent construct assessed at pretest. The difference between the latent means on the outcome variable in the treatment and control groups is compared after adjustment for the latent construct(s) represented by the covariates (Sörbom, 1979). Aiken, Stein, and Bentler (1994) illustrated the application of this approach to the nonequivalent control group design in their comparison of the effectiveness of two drug treatment programs.

GAIN SCORE ANALYSIS. Kenny (1975; see also Huitema, 1980; Judd & Kenny, 1981) has proposed an alternative approach to the analysis of the nonequivalent control group design. In this approach, the pretest and posttest scores are first transformed so that the pretest and posttest variances are equated. This transformation is most easily accomplished by separate standardizations of the pretest and posttest data using the pretest mean and standard deviation (see Huitema, 1980, chapter 15 for details). The mean gain in the treatment group is then compared with the mean gain in the control group to provide an estimate of the treatment effect. This approach provides good estimates of the treatment effect when it can be assumed that the natural pattern of growth in the absence of treatment is linear and constant in the two groups or is of a fan

[15] Cook and Reichardt (1976) and Judd and Kenny (1981) have suggested conducting sensitivity analyses using several plausible estimates of the reliability to bracket the true effect.

spread pattern, in which the growth is linear, but occurs at a higher rate for participants having higher scores at pretest. Many other forms of growth (or decline) may not be not well modeled by this procedure.

ECONOMETRIC SELECTION MODELS. A variety of selection models originally developed in econometrics have been applied to the nonequivalent control group design (e.g., Barnow, Cain, & Goldberger, 1980; Heckman, 1979, 1989, 1990; Muthén & Jöreskog, 1983). In these models, two separate regression equations are estimated. The first (selection) equation uses measured variables to predict the assignment of the participant to the treatment versus the control group and yields a probability for each participant of assignment to the treatment group. The second equation uses this selection probability, the treatment actually received by the participant, and other measured covariates to estimate the treatment effect. In practice, both equations are estimated simultaneously.

Selection models can be shown to provide highly accurate estimates of treatment effects when their assumptions are met. These models presume that highly reliable measures of all variables related to selection are included in the equation and that the sample size is relatively large. More critical, the results of these models are highly sensitive to violations of several additional statistical assumptions (e.g., the inclusion of a measured variable that affects selection, but is independent of the outcome measure) that must be made for proper estimation. Simulation studies by Stolzenberg and Relles (1990) and Virdin (1993) showed that selection models produced much poorer estimates than simpler techniques, such as analysis of covariance and gain score analysis when these assumptions are violated (see also Imbens & Rubin, 1994). Lalonde (1986) showed large discrepancies between causal effect estimates obtained from randomized experiments and econometric selection models.

**MATCHING: TRADITIONAL AND MODERN PROCEDURES.** In the simplest form of matching that is traditionally considered, each case in the treatment group is paired with a single case in the control group using a single measured covariate. To illustrate, consider a simplified example comparing two small school classrooms ($n_A = 12$; $n_B = 13$), one of which is to be given a new instructional treatment and one of which is to be given standard instruction (control group). Table 3.5 presents hypothetical data for this illustration in which the pretest (e.g., IQ) data have been ordered from low to high within each group. We note that adequate matches are available for 10 pairs of students. On the positive side, the mean difference on the pretest for the 10 matched pairs is considerably smaller ($\bar{X}_A - \bar{X}_B = 0.1$) than for the full, unmatched classes ($\bar{X}_A - \bar{X}_B = 11$). On the negative side, adequate matches are not available for all participants. The participants with the two highest and three lowest pretest scores must be dropped from the analysis; consequently, generalization to the full population of students will be limited.

The strategy of matching as an approach to equating the treatment and control groups at pretest in the nonequivalent control group design has been viewed differently by social psychologists and statisticians. Social psychological methodologists (e.g., Crano & Brewer, 1986; Judd, Smith, & Kidder, 1991) have focused on simple, traditional matching procedures, often arguing that matching is an especially flawed method of equating groups. They point to (a) the practical difficulty of finding good matches, particularly when the two groups must be matched on multiple covariates; (b) the frequent necessity of dropping participants from the analysis (illustrated in Table 3.5) because adequate matches cannot be found; and (c) the problem of regression to the mean, which we consider below.

Regression to the mean (see also Brewer, this volume, Ch. 1; Campbell & Kenny, 1999, Ch. 4) may occur under any one of three conditions. (a) Different measures are used to match participants at pretest (e.g., IQ) and to assess the outcome of the treatment at posttest (e.g., reading achievement scores). (b) The measures used to match participants do not have perfect internal consistency (see John & Benet-Martinez, this volume, Ch. 13). (c) The test–retest correlations between the pretest and posttest measure are less than 1.0. When disparate groups (e.g., children attending a public vs. a private school) are compared in nonequivalent control group designs, these groups are selected from two populations that may have different means on the covariates (e.g., the private school children may have a higher IQ and family income). Under these conditions, more children who are below the mean of their school on the measured covariate (e.g., IQ) will be systematically selected from the advantaged private school group, whereas more children who are above the mean of their school on the measured covariate will be systematically selected for matching from the disadvantaged public school group. In the case of matching on the basis of a single measured covariate, the combination of unreliability and systematic selection leads to the result that matching will underadjust for the initial

**TABLE 3.5. Illustration of Simple Matching of Two Classrooms on Pretest IQ Scores**

| Pair | Classroom A | Classroom B |
|---|---|---|
| | 150 | |
| | 130 | |
| 1 | 125 | 128 |
| 2 | 120 | 119 |
| 3 | 118 | 119 |
| 4 | 117 | 116 |
| 5 | 115 | 116 |
| 6 | 110 | 112 |
| 7 | 108 | 106 |
| 8 | 103 | 102 |
| 9 | 100 | 99 |
| 10 | 99 | 97 |
| | | 92 |
| | | 85 |
| | | 80 |

*Note.* Scores are ordered within classes and represent the pretest IQ measures of the students. Pairs of students on the same line represent matched pairs. Two students in Classroom A and three students in Classroom B have no matched pair and are thrown out of the design. The mean IQ for all 12 students in Classroom A is 116; the mean IQ for all 13 students in Classroom B is 105. For the 10 matched pairs, the mean IQ is 111.5 in Classroom A and 111.4 in Classroom B.

true differences between the means of the two groups on the matching variable. This is exactly the same problem we noted earlier with using analysis of covariance (without correction for unreliability) to adjust for initial differences between groups.

Statisticians have been more enthusiastic about the possibilities of matching and have developed techniques that can overcome several of the limitations noted above. Of special importance is Rosenbaum and Rubin's (1983, 1984) development of the concept of matching on the propensity score (see Rubin & Thomas, 1996). As in econometric selection models, the propensity score takes seriously the idea of attempting to develop a model of selection into the treatment and control groups. In this approach, a logistic regression equation is developed in which all of the available covariates are used to predict group assignment (treatment vs. control). For example, Rosenbaum and Rubin (1984) began with 74 available covariates in their comparison of the effectiveness of two treatments for heart disease in a study of 1,500 patients. Based on this equation, they estimated a propensity score for each participant that represents the probability that the specific participant would be assigned to the treatment group, given these background variables. These propensity scores are then used as the basis for matching participants in the treatment and control groups.

Rosenbaum and Rubin (1983, 1984) have shown that matching on propensity scores minimizes pretest differences between the treatment and control groups across the full set of measured variables. Rosenbaum (1986) also pointed out that matching adjusts for any functional form of relationship between the propensity score and the outcome measure, whereas the typical analysis of covariance only adjusts for linear relationships. Many of the earlier problems in finding adequate matches in the treatment and control groups can also be reduced through the use of modern computer search algorithms that identify the optimal pairing of participants[16] (Rosenbaum, 1995). Further, if matching on propensity scores is supplemented with other analyses, a variety of other potential issues can be probed (see Rosenbaum, 1986, 1987, 1995; Rubin & Thomas, 1996). The two following sets of supplementary analyses are of particular interest.

1. Although matching on propensity scores optimally balances the treatment and control groups across the set of covariates, it may not optimally match the two groups on variables of particular theoretical or empirical importance (e.g., the pretest measure of the outcome variable of interest). The treatment and control groups on each of the covariates may be compared using $t$-tests. If a researcher believes that certain covariates are especially critical and they do not appear to be well equated, the researcher may also perform analysis of covariance (or analysis of covariance with correction for unreliability to address issues of measurement error) to provide more precise adjustment for the effects of baseline differences in these specific variables.
2. A number of sensitivity analyses have been developed for matched group designs. These sensitivity analyses give an indication of the magnitude of bias

[16] The failure to find adequate matches highlights the limitations of generalization of the findings. Model-based statistical adjustments such as analysis of covariance imply that generalization across the full theoretical range of the covariate is possible. However, in regions where there are sparse data in one of the groups, such model-based generalization can represent a risky extrapolation of the findings (Cochran, 1957; see also Cook, 1993).

due to hidden variables that would be necessary to eliminate the treatment effect. As a simple example, Rosenbaum (1987) suggested calculating the standardized effect size $d$ for the covariate (not including the pretest) that shows the largest (unadjusted) difference at pretest – this estimate is used as a rough estimate of the magnitude of an important hidden variable. The correlation of the pretest and the posttest on the outcome variable of interest ($\rho_{12}$) is taken as the estimate of the maximum possible value of the relationship of the hidden variable with the outcome variable of interest. The product $d\rho_{12}$ is then used as a reasonable estimate of the amount that an important hidden variable not assessed at pretest could potentially reduce the estimate of the causal effect. If the difference between the treatment and control groups on the outcome variable can be reduced by this (or even a larger) amount and still attain statistical significance, then the estimate of the causal effect is taken as an indication that the treatment effect is likely to be robust to the effects of hidden variables. Rosenbaum (1991a, 1991b, 1994, 1995) also developed procedures for conducting sensitivity analyses using matched designs for cases involving a dichotomous outcome variable and for cases involving multiple continuous outcome variables. He has also developed procedures based on rank statistics that are less dependent on assuming that a specific statistical model (e.g., linear regression) is correct.[17]

In summary, the use of propensity scores overcomes many of the objections that were directed at earlier and simpler versions of matching. Particularly when supplemented by additional analyses to address specific issues (quality of matching on key covariates; unreliability on these key variables) and sensitivity analyses to address the likely impact of hidden variables, matching on the propensity score often provides distinct advantages over the sole use of statistical adjustment techniques. Belin et al. (1995) and Rosenbaum (1986) presented empirical illustrations of how the use of propensity scores combined with careful supplementary analyses can strengthen the conclusions that can be reached from the nonequivalent control group design. Recent meta-analytic studies in several areas of applied research have also provided an optimistic assessment of the success of careful matching procedures: These studies found no detectable difference in either the mean or variance of estimates of standardized treatment effects from randomized experiments and matched group designs (Heinsman & Shadish, 1996; Shadish & Ragsdale, 1996). Nonetheless, a strong caveat remains: The fundamental condition of strong ignorability that is necessary for the causal interpretation of treatment effects in the nonequivalent control group design can be probed, but never definitively established. Thus, there is always a degree of uncertainty associated with estimates of causal effects on the basis of this design.

[17] Rosenbaum (1986, 1987, 1995) also provided illustrations of a third type of supplementary analysis. He showed how the combination of propensity score matching and substantive theory in an area can be used to make predictions that provide checks on the assumption of strong ignorability. Because his examples depend on theory in substantive areas other than social psychology, they will not be discussed here.

**THREATS TO INTERNAL VALIDITY.** In terms of Campbell's approach, the basic nonequivalent control group design is particularly susceptible to four threats to internal validity (Cook & Campbell, 1979). The strength of the design is that the inclusion of the control group rules out basic threats, such as maturation, that occur equally in the treatment and control groups. The weakness of the basic design is that it does not rule out cases in which these threats operate differently in the treatment and control groups.

MATURATION. The treatment and control groups may differ in the rate at which the participants are growing or declining on the outcome variable prior to treatment. To illustrate, consider a nonequivalent control group design in which the Evans et al. (1981) smoking prevention program is given to all students in a suburban high school, whereas students in an inner city high school receive a control program unrelated to smoking. An example of the threat of differential maturation would occur if the number of cigarettes smoked per day in an urban high school was increasing at a faster rate than in the suburban high school in the absence of intervention.

HISTORY. Some other event may occur to one of the two groups, but not the other, that may be expected to affect the outcome variable. For example, the media in the suburban site might independently start a series pointing out the dangers of teenage smoking. Alternatively, lower cost generic or contraband cigarettes might become available in the urban, but not in the suburban area.

STATISTICAL REGRESSION. Participants may be selected for the treatment or control group based on an unreliable or temporally instable measured variable.

Participants selected for the study in the inner city school may have on average been temporarily smoking fewer than their normal number of cigarettes per week (e.g., if several of the participants had just recovered from colds), whereas participants in the suburban school may have on average been smoking their normal number of cigarettes per day. Upon remeasurement, participants in each group would tend to report their typical level of smoking.

INSTRUMENTATION. Some aspect of the measuring instrument may change from pretest to posttest in only one of the two groups. This threat can take many forms, which can be manifested in such problems as differences in the factor structures (see Pitts, West, & Tein, 1996), the reliability of the scale, or the interval properties on the scale itself between the two groups (e.g., ceiling or floor effects). Local changes in record keeping practices or in the sensitivity of the measures can also produce this threat.

SUMMARY. In the example given above, the smoking prevention program was expected to reduce the number of cigarettes that were smoked by the average student. Given that the suburban high school was assigned to receive the smoking prevention program, any observed treatment effect is confounded with the specific forms of the threats of maturation, history, regression, and instrumentation presented above. Any observed difference between the means of the treatment and control group means may be due to the causal effect of treatment, the specific threats to internal validity, or both. However, note that if the smoking prevention intervention had been assigned to the inner city instead of the suburban school, none of the specific forms of threats to internal validity illustrated above would provide a plausible alternative explanation of the effects of the program. In this second case, the threats to internal validity bias the estimates of the causal effect, but in a direction opposite from that expected for true effects of the program. These specific threats may lead to estimates of the causal effect of treatment that are too low, thus lowering the statistical power of the test. However, these threats do not call the existence of the causal effect into question.

**DESIGN ENHANCEMENTS.** The basic nonequivalent control groups design may be strengthened by adding design and analysis components that specifically address the plausible threats to validity. The ideas here build on previous discussions of design enhancements for addressing threats to internal validity in other quasi-experimental designs. Reynolds and West (1987) presented extensive examples of the use of several of these techniques in the evaluation of the effectiveness of a sales campaign.

MULTIPLE CONTROL GROUPS. Typically, no control group can be identified that is comparable to the treatment group on all factors that could affect the outcome of the study. It is often possible, however, to identify multiple imperfect control groups, each of which can address some of the threats to validity. For example, Roos, Roos, and Henteleff (1978) compared the pre- and postoperation health of tonsillectomy patients with (a) a nonoperated control group matched on age and medical diagnosis and (b) nonoperated siblings who were within 5 years of the patient's age. The first control group roughly equates the medical history of the treatment participant; the second control group roughly equates the genetic predisposition, environment, and family influences of the treatment participant. Obtaining similar estimates of the causal effects across each of the comparison groups can help minimize the plausibility of alternative explanations of the obtained results. Rosenbaum (1987; see also associated discussion papers) discussed several methods of using the data from carefully chosen multiple control groups to reduce the likelihood that hidden variables may be responsible for the results.

NONEQUIVALENT DEPENDENT VARIABLES. Sometimes data can be collected on additional dependent variables or in different contexts on the same group of participants. Ideally, the measures selected should be conceptually related to the outcome variable of interest and should be affected by the same threats to internal validity as the primary dependent variable. However, the researcher would *not* expect these measures to be affected by the treatment. To illustrate, Roos et al. (1978) compared the tonsillectomy group and two control groups on their primary outcome variable, health insurance claims for respiratory illness, which they expected to decrease only in the treatment group. They also compared health insurance claims for other nonrespiratory illness that would not be expected to decrease following treatment. To the extent that the hypothesized pattern of results is obtained on the primary outcome variable, but not on the nonequivalent dependent variables, the interpretation of the results as a causal effect is strengthened.

MULTIPLE PRETESTS OVER TIME. The addition of multiple pretests on the same variable over time to the nonequivalent control group design can help rule out threats associated with differential rates of maturation

or regression to the mean in the two groups. In this design, the multiple pretests are used to estimate the rates of maturation in the two groups prior to treatment. These estimates can then be used to adjust the treatment effect under the assumption that the pattern of maturation within each group would not change during the study. For example, Reynolds and West (1987) used data from sales in treatment and comparison stores prior to the introduction of a sales campaign in treatment stores to adjust for potential (maturational) trends in sales levels. Recent applications of hierarchical linear models to repeated measures data (Bryk & Raudenbush, 1992; Willett & Sayers, 1994) have provided a number of promising and flexible models of adjusting the treatment effect based on information about each participant's growth rate and the mean growth rates within the treatment and control groups (see Pitts, 1999, for a detailed development of possible models).

**PATTERN OF RESULTS.** As we briefly noted previously, not all threats to internal validity are plausible given certain patterns of results. Cook and Campbell (1979) noted that the four threats to validity may increase or decrease in plausibility depending on the specific area of research and the pattern of results that is obtained. For example, cases of differential rates of maturation almost never involve cases in which only one group is growing or the two groups are growing in opposite directions; rather, they involve differential rates of either growth or decline in the two groups. Similarly, regression to the mean involves a disadvantaged group improving or an advantaged group declining on remeasurement relative to another selected group. It is implausible to expect that regression to the mean would lead a previously disadvantaged group to surpass a previously advantaged group.

Figure 3.6 draws on these ideas and helps identify the likely threats to internal validity that would be associated with several different possible outcome patterns. For example, applying the previous reasoning to the examples, outcomes B and E suggest that differential rates of maturation should be carefully considered as a potential threat to the internal validity of the results. In contrast, differential rates of maturation is a less likely threat given outcomes A, C, or particularly D.

Using the guidelines presented by Figure 3.6, researchers can identify those threats to internal validity that deserve special attention given the specific pattern of results obtained in their study. For example, if we consider outcome D, the threats of differential maturation rates, instrumentation, and regression are unlikely, leaving differences in history in the two groups as the threat to the internal validity toward which researchers should focus their attention. If examination of careful documentation of other possible influences on the treatment and control groups failed to identify plausible historical influences that differed in the two groups, the researchers could reach at least a tentative conclusion that their treatment led to a causal effect.

### OTHER ISSUES

STATISTICAL POWER. As noted previously, to the extent the condition of strong ignorability can be approached, the nonequivalent control group design theoretically becomes increasingly similar to the regression discontinuity design. Because in practice researchers cannot convincingly establish that the assumptions necessary for strong ignorability have been met, multiple analyses must often be conducted to probe assumptions and to rule out plausible threats to validity. Methodologists frequently suggest that researchers attempt to bracket the estimate of the treatment effect, sometimes reporting the mean (or median), largest, and smallest effect sizes across diverse analyses (see Reichardt & Gollob, 1997, for a discussion). This strategy increases the certainty of our causal inference to the extent that all of the analyses yield similar estimates of the treatment effect. However, this strategy also implies that if we wish to show that the bracket does not include a treatment effect of 0, our estimate of statistical power must be based on the statistical model that can be expected to lead to the greatest attenuation of the magnitude of the treatment effect. As a result, the power of the statistical tests in the nonequivalent control group design can be expected to be lower than for a randomized experiment.

**CAUSAL GENERALIZATION.** Nonequivalent control group designs are normally conducted with the units and settings of interest and are thus typically relatively high in external validity. Randomization can and should be implemented more frequently than it has been (Cook & Campbell, 1979; Rosenbaum, 1986; Shadish et al., in press), but many potential participants and organizational contexts will not accept this assignment procedure. Consequently, the nonequivalent control group design, ideally including the design enhancements discussed previously, is sometimes the only feasible option. For example, it would be difficult to get public and private schools to accept random assignment of students; furthermore, not all students (or their families) would be willing to participate in such an assignment process. Even if a random assignment process could be implemented, it is likely that the

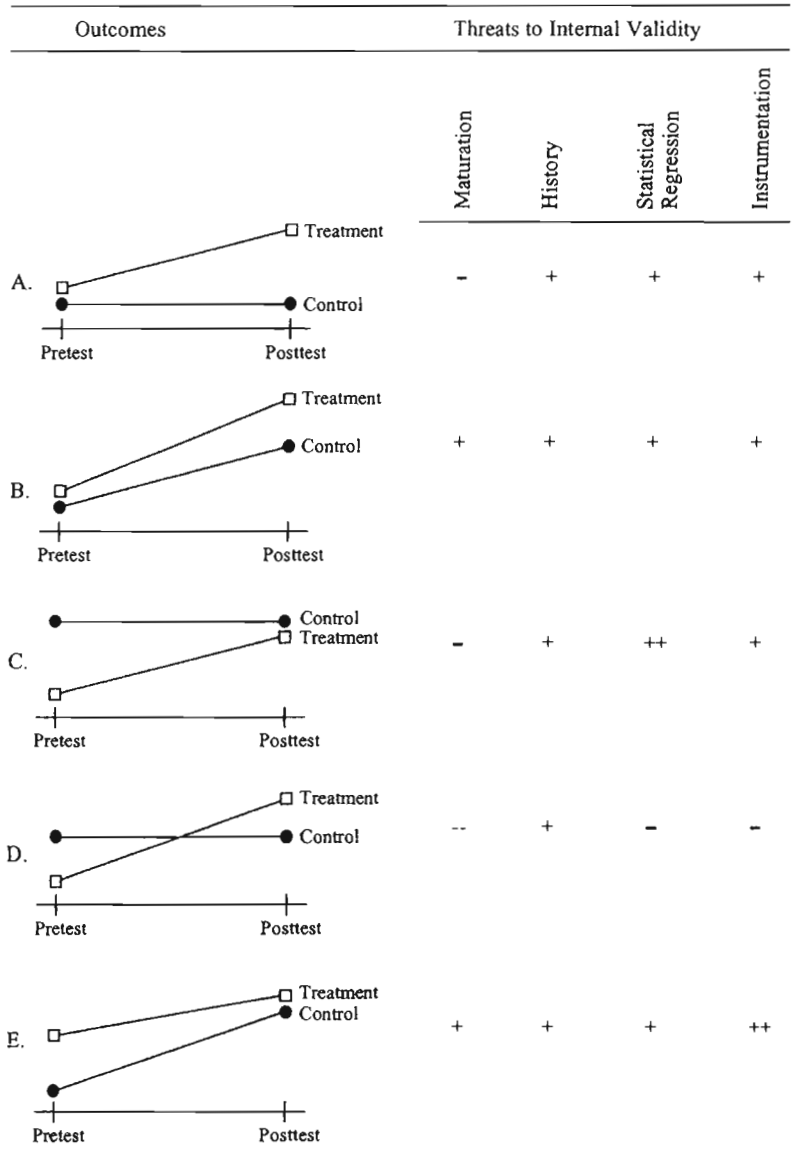



**Figure 3.6.** Interpretable outcomes associated with the nonequivalent control group design.

*Note.* ++ indicates a highly likely threat to the design when this outcome occurs.
+ indicates a likely threat to the design when this outcome occurs.
-- indicates a less likely threat to the design when this outcome occurs.
- indicates a very unlikely threat to the design when this outcome occurs.
Adapted from S. G. West, Beyond the laboratory experiment, chapter in P. Karoly (Ed.), *Measurement strategies in health psychology*, p. 227. Reprinted by permission of John Wiley and Sons.

participating families and schools would be highly atypical of the population of units and settings to which the researchers intend to generalize their results (West & Sagarin, 2000). In addition, treatments and observations in nonequivalent control group designs are often more similar to those that will be implemented in actual settings, further enhancing causal generalization to the UTOS of interest. The ability to generalize causal effects to the UTOS of interest is one of the primary strengths of the nonequivalent control group design; the inability to make precise statements about the magnitude (and sometimes direction) of causal effects is its primary weakness.

**CONCLUSION.** The basic nonequivalent control group design provides the least satisfactory of the three quasi-experimental alternatives to the randomized experiment we have considered. Such designs have traditionally been viewed as "very weak, easily misinterpreted, and difficult to analyze" (Huitema, 1980, p. 352). From the perspective of the RCM, there is always some uncertainty as to whether the conditions of strong ignorability and SUTVA have been met. From the perspective of Cook and Campbell (1979) and Shadish et al. (in press), the four threats of differential maturation rates, history, statistical regression, and instrumentation in the two groups represent the

primary threats to internal validity that must be ruled out. The internal validity of the design can potentially be enhanced by the inclusion of design features that address specific threats to internal validity, such as multiple control groups, nonequivalent dependent measures, and multiple pretests over time, (Reynolds & West, 1987). Data may sometimes be collected that permit the researcher either to adequately model the selection process or match groups on propensity scores. The use of multiple statistical adjustment procedures that have different assumptions, but which converge on similar estimates of the causal effect can also produce increased confidence in the results. In general, strong treatment effects that can be shown through sensitivity analyses to be robust to the potential influence of hidden variables coupled with careful consideration of the remaining threats to internal validity can sometimes overcome most of the limitations of this design.

## SOME FINAL OBSERVATIONS

This chapter has provided an introduction to experimental and quasi-experimental designs that are useful in field settings. In contrast to the laboratory research methods, in which social psychologists have traditionally been trained, modern field research methodology reflects the more complex and less certain real-world settings in which it has been applied. As we have seen, even the randomized experiment may not definitively rule out all threats to internal validity – problems such as attrition, treatment noncompliance, and atypical responses of control groups occur with some regularity. Weaker quasi-experimental designs, such as the nonequivalent control group design, rule out a far smaller set of the threats to internal validity. Consequently, social psychologists working in field settings must carefully articulate the threats to internal validity and use specific procedures in an attempt to minimize the likelihood that a specific threat to internal validity could account for the results. These procedures include adding specific design features (e.g., multiple control groups), additional measurements (e.g., time series data), and statistical adjustments (see also Reichardt & Mark, 1997). The use of these procedures can in some cases produce strong designs whose internal validity approaches that of a randomized experiment. In other cases, one or more threats to internal validity will remain plausible despite the investigator's best efforts.

Because of the complexity and uncertainty associated with research conducted in the field, it is important for researchers to acknowledge publicly the known limitations of their findings and to make their data available for analysis by other researchers (Cook, 1983; Houts et al., 1986; Sechrest, West, Phillips, Redner, & Yeaton, 1979). Public criticism of research provides an important mechanism through which biases and threats to internal validity can be identified and their likely effects, if any, on the results of the study can be assessed. Additional studies, with different strengths and weaknesses, can then be conducted to help bracket the true causal effect. Although considerable uncertainty may be associated with the results of any single study, a consistent finding in the research literature can greatly increase confidence in the robustness of the causal effect (Shadish et al., in press).

Reflecting recent practice, the majority of the research examples discussed in this chapter have been applied in nature. The development of the methods discussed in this chapter has provided a strong basis for applied social psychologists to make causal inferences about the effects of treatment programs delivered in the settings of interest in the field. At the same time, social psychologists interested in basic research are beginning to develop new areas of substantive interest. Among these areas are the influence of culture (e.g., collectivist vs. individualist), major life stressors, long-term relationships, and new aspects of the self (e.g., generativity). Methodological issues that arise in these areas, such as selection of participants into conditions, attrition, and growth or decline over time require that researchers consider many of the specific design, measurement, and statistical techniques developed to address these problems. Interest in problems such as coping with the stress of physical illness often require that researchers collect data in settings outside the laboratory. And interest in problems like long-term relationships and generativity will often require that researchers study relevant samples of participants over extended time periods. These substantive developments suggest that the traditional modal study in social psychology identified in West et al.'s (1992) review – a randomized experiment, conducted in the laboratory, lasting no more than 1 hour, and using undergraduate students as participants – may no longer represent the design of choice in many of these emerging research areas. These substantive developments call for new variants of the laboratory experiment that incorporate some of the features of modern field experiments and quasi-experiments discussed in this chapter. These developments also may portend a possible return to modern improved versions of research methods more commonly used in some

previous eras in the history of social psychology – experiments, quasi-experiments, and other studies testing basic social psychological principles in field contexts. Such potential developments would help social psychology broaden the Units, Treatments, Observations, and Settings represented in its research base, providing a stronger basis for causal generalization. They would supplement the demonstrated strengths and complement the weaknesses of traditional laboratory experiments. Such designs hold the promise of a more balanced mix of methodological approaches to basic issues in social psychology, in which researchers could make legitimate claims for both the internal validity and the causal generalization of their effects.

## REFERENCES


Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology, 62,* 488–499.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review, 22,* 207–244.

Aiken, L. S., West, S. G., Woodward, C. K., Reno, R. R., & Reynolds, K. D. (1994). Increasing screening mammography in asymptomatic women: Evaluation of a second generation, theory-based program. *Health Psychology, 13,* 526–538.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with commentary). *Journal of the American Statistical Association, 91,* 444–472.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–278). Mahwah, NJ: Erlbaum.

Aronson, E. (1966). Avoidance of inter-subject communication. *Psychological Reports, 19,* 238.

Aronson, E., Wilson, T. D., & Brewer (1998). Experimentation in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 1, pp. 99–142). Boston: McGraw-Hill.

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice Hall.

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6,* 267–285.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon.

Barnow, L. S. (1973). The effects of Head Start and socioeconomic status on cognitive development of disadvantaged students (Doctoral dissertation, University of Wisconsin, Madison, 1974). *Dissertation Abstracts International, 34,* 6191A.

Barnow, L. S., Cain, G. G., & Goldberger, A. S. (1980). Issues in the analysis of selection bias. In E. S. Stromsdorfer & G. Farkas (Eds.), *Evaluation studies review annual* (Vol. 5, pp. 43–59). Beverly Hills, CA: Sage.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Belin, T. R., Elashoff, R. M., Leung, K.-M., Nisenbaum, R., Bastani, R., Nasseri, K., & Maxwell, A. (1995). Combining information from multiple sources in the analysis of a non-equivalent control group design. In C. Gatsonis, J. S. Hodges, R. E. Kass, & N. D. Singpurwalla (Eds.), *Case studies in Bayesian statistics* (Vol. 2, pp. 241–260). New York: Springer-Verlag.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Woodward, J. A. (1978). A Head Start re-evaluation: Positive effects are not yet demonstrable. *Evaluation Quarterly, 2,* 493–510.

Berk, R. A. (1988). Causal inference for sociological data. In N. J. Smeltzer (Ed.), *Handbook of sociology* (pp. 155–172). Newbury Park, CA: Sage.

Berkowitz, L. (1993). *Aggression: Its causes, consequences, and control.* New York: McGraw-Hill.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep. *American Psychologist, 37,* 245–257.

Bickman, L., & Henchy, T. (Eds.). (1972). *Beyond the laboratory: Field research in social psychology.* New York: McGraw-Hill.

Biglan, A., Hood, D., Borzovsky, P., Ochs, L., Ary, D., & Black, C. (1991). Subject attrition in prevention research. In C. G. Luekefeld & W. Bukoski (Eds.), *Drug abuse prevention intervention research: Methodological issues* (pp. 213–234). Washington, DC: NIDA Research Monograph # 107.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review, 8,* 225–246.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Borenstein, M., Cohen, J., & Rothstein, H. (1997). *Confidence intervals, effect size, and power* [Computer program]. Mahwah, NJ: Erlbaum.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation.* Thousand Oaks, CA: Sage.

Boruch, R. F., McSweeny, A. J., & Soderstrom, E. J. (1978). Randomized field experiments for program planning, development, and evaluation. *Evaluation Quarterly, 2,* 655–695.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model building and response surfaces*. New York: Wiley.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control* (3rd ed.). San Francisco: Holden-Day.

Braucht, G. N., & Reichardt, C. S. (1993). A computerized approach to trickle-process, random assignment. *Evaluation Review, 17*, 79–90.

Bryan, A. D., Aiken, L. S., & West, S. G. (1996). Increasing condom use: Evaluation of a theory-based intervention to prevent sexually transmitted disease in young women. *Health Psychology, 15*, 371–382.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In D. Berliner (Ed.), *Review of research in education* (Vol. 8, pp. 158–223). Washington, DC: American Educational Research Association.

Burstein, L. (1985). Units of analysis. In *International encyclopedia of education* (pp. 5368–5375). Oxford, England: Pergamon.

Burstein, L., Kim, K.-S., & Delandshire, G. (1989). Multilevel investigations of systematically varying slopes: Issues, alternatives, and consequences. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 233–276). San Diego, CA: Academic Press.

Cahan, S., & Davis, D. (1987). A between-grade-levels approach to the investigation of the absolute effects of schooling on achievement. *American Educational Research Journal, 24*, 1–12.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297–312.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24* , 409–429.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (Vol. 31, pp. 67–78). San Francisco: Jossey-Bass.

Campbell, D. T., & Boruch, R. F. (1975). Making the case for randomized assignment by considering the alternatives: Six ways in which quasi-experimental evaluations tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experience: Some critical issues in assessing social programs* (pp. 195–296). New York: Academic Press.

Campbell, D. T., & Erlbacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Compensatory education: A national debate. Volume. 3: Disadvantaged child* (pp. 185–225). New York: Brunner/Mazel.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Canner, P. (1984). How much data should be collected in a clinical trial? *Statistics in Medicine, 3*, 423–432.

Canner, P. (1991). Covariate adjustment of treatment effects in clinical trials. *Controlled Clinical Trials, 12*, 359–366.

Cappelleri, J. C. (1990, October). *Power analysis of regression-discontinuity designs*. Paper presented at the annual meeting of the American Evaluation Association, Washington, DC.

Cappelleri, J. C., Darlington, R. B., & Trochim, W. M. K. (1994). Power analysis of cutoff-based randomized clinical trials. *Evaluation Review, 18*, 141–152.

Chassin, L., Barrera, M., Jr., Bech, K., & Kossak-Fuller, J. (1992). Recruiting a community sample of adolescent children of alcoholics: A comparison of three subject sources. *Journal of Studies on Alcohol, 53*, 316–319.

Chatfield, C. (1996). *The analysis of time series: An introduction* (5th ed.). London: Chapman & Hall.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*, 1015–1026.

Cicarelli, V. G., Cooper, W. H., & Granger, R. L. (1969). *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development*. Athens, OH: Ohio University and Westinghouse Learning Corporation.

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics, 13*, 261–281.

Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A, 128*, 134–155.

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs* (6th ed.). New York: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round, $p < .05$. *American Psychologist, 49*, 997–1003.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (in press). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*.

Coleman, J., Hoffer, T., & Kilgore, S. (1982). Cognitive outcomes in public and private schools. *Sociology of Education, 55*, 65–76.

Conner, R. F. (1977). Selecting a control group: An analysis of the randomization process in twelve social reform programs. *Evaluation Quarterly, 1*, 195–244.

Cook, R. D., & Weisberg, S. (1994). *An introduction to regression graphics*. New York: Wiley.

Cook, T. D. (1983). Quasi-experimentation: Its ontology, epistemology, and methodology. In G. Morgan (Ed.), *Beyond method: Strategies for social research*. Beverly Hills, CA: Sage.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. B. Sechrest & A. G. Scott (Eds.). *New directions for program evaluation*, (Number 57, 39–81). San Francisco: Jossey-Bass.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.

Cook, T. D., & Reichardt, C. S. (1976). Statistical analysis of data from the nonequivalent control group design: A guide to some current literature. *Evaluation, 3*, 136–138.

Crano, W. D., & Brewer, M. B. (1986). *Principles and methods of social research* (2nd ed.). Boston: Allyn & Bacon.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis*. Occasional Paper, Stanford Evaluation Consortium, Stanford, CA.

Cronbach, L. J. (1982). *Designing evaluations of social and educational programs*. San Francisco: Jossey-Bass.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61*, 966–974.

Daniel, C., & Wood, F. S. (1980). *Fitting equations to data* (2nd ed.). New York: Wiley.

Dennis, M. L., Lennox, R. D., & Williams, R. (1997). Practical power analysis. In K. Bryant, M. Windle, & S. G. West (Eds.), *The Science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 367–404). Washington, DC: American Psychological Association.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics, 20*, 115–147.

Durlak, J. A., & Wells, A. M. (1997). Primary prevention mental health programs for children and adolescents: A meta-analytic review. *American Journal of Community Psychology, 25*, 115–152.

Efron, B., & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials (with discussion). *Journal of the American Statistical Association, 86*, 9–26.

Erlebacher, A. (1977). Design and analysis of experiments contrasting the within- and between-subjects manipulation of the independent variable. *Psychological Bulletin, 84*, 212–219.

Evans, R. I., Rozelle, R. M., Maxwell, S. E., Raines, B. E., Dill, C. A., Guthrie, T. J., Henderson, A. H., & Hill, P. C. (1981). Social modeling films to deter smoking in adolescents: Results of a three-year field investigation. *Journal of Applied Psychology, 66*, 399–414.

Fisher, R. A. (1935). *The design of experiments*. London: Oliver & Boyd.

Flay, B. R. (1986). Psychosocial approaches to smoking prevention: A review of findings. *Health Psychology, 4*, 449–488.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (1996). *Design and analysis of single case research*. Mahwah, NJ: Erlbaum.

Funder, D. C. (1992). Psychology from the other side of the line: Editorial processes and publication trends at *JPSP*. *Personality and Social Psychology Bulletin, 18*, 493–497.

Gilbert, J. P., Light, R. J., & Mosteller, F. (1975). Assessing social innovations: An empirical base for policy. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs*. (pp. 39–193) New York: Academic Press.

Glass, G. V. (1988). Quasi-experiments: The case of interrupted time series. In R. M. Jaeger (Ed.), *Complementary methods for research in education* (pp. 445–464). Washington, DC: American Educational Research Association.

Goldberger, A. S. (1972, April). *Selection bias in evaluating treatment effects: Some formal illustrations* (Discussion paper 123–72). Madison: University of Wisconsin, Institute for Research on Poverty.

Greenland, S., & Morgenstern, H. (1989). Ecological bias, confounding, and effect modification. *International Journal of Epidemiology, 18*, 269–274.

Greenwald, A. G. (1976). Within-subjects design: To use or not to use. *Psychological Bulletin, 83*, 314–320.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review, 93*, 216–229.

Gregson, R. A. (1987). The time-series analysis of self-reported headache symptoms. *Behavior Change, 4*(2), 6–13.

Hansen, W. B., & Collins, L. M. (1994). Seven ways to increase power without increasing N. In L. M. Collins & L. A. Seitz (Eds.), *Advances in data analysis for prevention intervention research* (pp. 184–195). Rockville, MD: NIDA Research Monograph 142. NIH Publication No. 94-3599.

Hansen, W. B., Collins, L. M., Malotte, C. K, Johnson, C. A., & Fielding, J. E. (1985). Attrition in prevention research. *Journal of Behavioral Medicine, 8*, 261–275.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.).(1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman & Hall.

Heckman, J. J. (1979). Sample bias as a specification error. *Econometrica, 46*, 153–162.

Heckman, J. J. (1989). Causal inference and nonrandom samples. *Journal of Educational Statistics, 14*, 159–168.

Heckman, J. J. (1990). Varieties of selection bias. *American Economic Review, 80*, 313–318.

Hedecker, D., Gibbons, R. D., & Flay, B. R. (1994). Random effects regression models for clustered data with an example from smoking prevention research. *Journal of Consulting and Clinical Psychology, 62*, 757–765.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments. *Psychological Methods, 1,* 154–169.

Hennigan, K. M., del Rosario, M. L., Heath, L., Cook, T. D., Wharton, J. L., & Calder, B. J. (1982). Impact of the introduction of television on crime in the United States: Empirical findings and theoretical implications. *Journal of Personality and Social Psychology, 42,* 461–477.

Higginbotham, H. N., West, S. G., & Forsyth, D. R. (1988). *Psychotherapy and behavior change: Social, cultural, and methodological perspectives.* New York: Pergamon.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association, 81,* 945–970.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models (with discussion). In C. Clogg (Ed.), *Sociological methodology 1988* (pp. 449–493). Washington, DC: American Sociological Association.

Holland, P. W. (1989). Discussion of Aitkin's and Longford's papers. In R. D. Bock (Ed.),*Multilevel analysis of educational data* (pp. 311–317). San Diego, CA: Academic.

Horne, G. P., Yang, M. C. K., & Ware, W. B. (1982). Time series analysis for single subject designs. *Psychological Bulletin, 91,* 178–189.

Houts, A. C., Cook, T. D., & Shadish, W. R. (1986). The person-situation debate: A critical multiplist perspective. *Journal of Personality, 54,* 52–105.

Huitema, B. E. (1980). *The analysis of covariance and alternatives.* New York: Wiley.

Hunter, J. E. (1996, August). Needed: A ban on the significance test. In P. E. Shrout (chair), Symposium. Significance tests – should they be banned from APA journals. American Psychological Association, Toronto, Ontario, Canada.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Imbens, G. W., & Rubin, D. B. (1994). *On the fragility of instrumental variables estimators.* Discussion paper # 1675, unpublished manuscript, Harvard University, Institute of Economic Research.

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics , 25,* 305–327.

Ito, T. A., Miller, N., & Pollock, V. E. (1996). Alcohol and aggression: A meta-analysis on the moderating effects of inhibitory cues, triggering events, and self-focused attention. *Psychological Bulletin, 120,* 60–82.

Jöreskog, K. G., & Sörbom, D. (1993). *Lisrel 8: User's reference guide.* Chicago: Scientific Software.

Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions.* New York: Cambridge University Press.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). *Research methods in social relations* (6th ed.). Fort Worth, TX: Harcourt Brace Jovanovich.

Jurs, S. G., & Glass, G. V. (1971). The effect of experimental mortality on the internal and external validity of the randomized comparative experiment. *Journal of Experimental Education, 40,* 62–66.

Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin, 82,* 345–362.

Kenny, D. A., & Judd, C. M. (1986). The consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 82,* 345–362.

Keren, G. (1993). Between-or within-subjects design: A methodological dilemma. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 257–272). Hillsdale, NJ: Erlbaum.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brook/Cole.

Kish, L. (1987). *Statistical designs for research.* New York: Wiley.

Kopans, D. B. (1994). Screening for breast cancer and mortality reduction among women 40–49 years of age. *Cancer, 74* (Suppl), 311–322.

Kratochwill, T. R., & Levin, J. R. (Eds.). (1992). *Single-case research design and analysis: New directions for psychology and education.* Hillsdale, NJ: Erlbaum.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling.* London: Sage.

Lalonde, R. (1986). Evaluating the econometric evaluations of training programs. *American Economic Review, 76,* 604–620.

Larsen, R. J. (1989). A process approach to personality psychology: Utilizing time as a facet of data. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 177–193). New York: Springer-Verlag.

Lee, Y., Ellenberg, J., Hirtz, D., & Nelson, K. (1991). Analysis of clinical trials by treatment actually received: Is it really an option? *Statistics in Medicine, 10,* 1595–1605.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday events. *American Psychologist, 43,* 431–442.

Linn, R. L., & Werts, C. E. (1973). Errors of inference due to errors of movement. *Educational and Psychological Measurement, 33,* 531–545.

Lipsey, M. W. (1993). Theory as method: Small theories of treatments. In L. B. Sechrest & A. G. Scott (Eds.), *New directions in program evaluation.* (Number 57, pp. 5–38). San Francisco: Jossey-Bass.

Lipsey, M. W. (1997). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. Rog (Eds.), *Handbook of applied social research methods* (pp. 39–68). Thousand Oaks, CA: Sage.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Little, R. J. A., & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39–76). New York: Plenum.

Little, R. J., & Yau, L. H. Y (1998). Statistical techniques for analyzing data from preventive trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3,* 147–159.

MacKinnon, D. P. (1994). Analysis of mediating variables in prevention and intervention research. In A. Cezares & L. Beatty (Eds.), *Scientific methods for prevention intervention research.* Rockville, MD: National Institute on Drug Abuse.

Magidson, J. (1977). Toward a causal modeling approach to adjusting for pre-existing differences in the nonequivalent group situation: A general alternative to ANCOVA. *Evaluation Quarterly, 1,* 399–420.

Mallar, C. D., & Thornton, C. V. D. (1978). Transitional aid for released prisoners: Evidence from the life experiment. In T. D. Cook, M. L. Del Rosario, K. M. Hennigan, M. M. Mark, & W. M. K. Trochim (Eds.), *Evaluation studies review annual* (Vol. 3, pp. 498–517). Beverly Hills, CA: Sage.

Marcantonio, R. J., & Cook, T. D. (1994). Convincing quasi-experiments: The interrupted time series and regression-discontinuity designs. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 133–254). San Francisco: Jossey-Bass.

Mark, M. M., & Mellor, S. (1991). Effect of self-relevance of an event on hindsight bias: The foreseeability of a layoff. *Journal of Applied Psychology, 76,* 569–577.

Martin, S. E., Annan, S., & Forst, B. (1993). The special deterrent effects of a jail sanction on first-time drunk drivers: A quasi-experimental study. *Accident Analysis and Prevention, 25,* 561–568.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective.* Pacific Grove, CA: Brooks/Cole.

McCleary, R., & Hay, R. A. (1980). *Applied time series analysis.* Beverly Hills, CA: Sage.

McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 1, pp. 192–229). New York: Academic Press.

McGuire, W. J. (1983). A contextualist theory of knowledge: Its implications for innovation and research in psychological research. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 16, pp 1–47). New York: Academic Press.

McSweeny, A. J. (1978). The effects of response cost on the behavior of a million persons. *Journal of Applied Behavior Analysis, 11,* 47–51.

Meier, P. (1985). The biggest public health experiment ever: The 1954 field trial of the Salk polio vaccine. In J. M. Tanur, F. Mosteller, W. H. Kruskal, R. F. Link, R. S. Pieters, & G. R. Rising (Eds.), *Statistics: A guide to the unknown* (2nd Ed.) (pp. 3–15). Monterey, CA: Wadsworth & Brooks/Cole.

Meier, P. (1991). Comment. *Journal of the American Statistical Association, 86,* 19–22.

Mohr, L. B. (1988). *Impact analysis for program evaluation.* New York: Basic Books.

Murname, R. J., Newstead, S., & Olson, F. J. (1985). Comparing public and private schools: The puzzling role of selectivity bias. *Journal of Business and Economic Statistics, 3,* 23–35.

Muthén, B., & Jöreskog, K. G. (1983). Selectivity problems in quasi-experimental studies. *Evaluation Review, 7,* 139–174.

Muthén, B., Kaplan, M., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52,* 431–462.

Pitts, S. C. (1999). *The use of latent growth models to estimate treatment effects in longitudinal experiments.* Unpublished doctoral dissertation, Arizona State University.

Pitts, S. C., West, S. G., & Tein, J.-Y. (1996). Longitudinal measurement models in evaluation research: Examining stability and change. *Evaluation and Program Planning, 19,* 333–350.

Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. D. Cook and D. T. Campbell, *Quasi-experimentation: Design and analysis issues for field settings* (pp. 147–205). Boston: Houghton-Mifflin.

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 259–284). Mahwah, NJ: Erlbaum.

Reichardt, C. S., & Mark, M. M. (1997). Quasi-experimentation. In L. Bickman & D. Rog (Eds.), *Handbook of applied social research methods* (pp. 193–228). Thousand Oaks, CA: Sage.

Reichardt, C. S., Minton, B. A., & Schellenger, J. D. (1986). *The analysis of covariance (ANCOVA) and the assessment of treatment effects.* Unpublished manuscript, University of Denver.

Reichardt, C. S., Trochim, W. M. K., & Cappelleri, J. C. (1995). Reports of the death of regression-discontinuity analysis are greatly exaggerated. *Evaluation Review, 19,* 39–63.

Reis, H. T., & Stiller, J. (1992). Publication trends in *JPSP*: A three-decade review. *Personality and Social Psychology Bulletin, 18,* 465–472.

Reynolds, K. D., & West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review, 11,* 691–714.

Ribisl, K. M., Watlon, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. A., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning, 19,* 1–25.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357.

Roos, L. L., Jr., Roos, N. P., & Henteleff, P. D. (1978). Assessing the impact of tonsillectomies. *Medical Care, 16,* 502–518.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79,* 41–48.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics, 11,* 207–224.

Rosenbaum, P. R. (1987). The role of a second control group in an observational study (with discussion). *Statistical Science, 2,* 292–316.

Rosenbaum, P. R. (1991a). Discussing hidden bias in observational studies. *Annals of Internal Medicine, 115,* 901–905.

Rosenbaum, P. R. (1991b). Sensitivity analysis for matched case-control studies. *Biometrics, 47,* 87–100.

Rosenbaum, P. R. (1994). Coherence in observational studies. *Biometrics, 50,* 368–374.

Rosenbaum, P. R. (1995). *Observational studies.* New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.

Rosenbaum, P. R., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516–524.

Rosenthal, R., & Rubin, D. (1980). Comparing within- and between-subjects studies. *Sociological Methods and Research, 9,* 127–136.

Ross, H. L., Campbell, D. T., & Glass, G. V. (1970). Determining the social effects of a legal reform: The British "breathalyser" crackdown of 1967. *American Behavioral Scientist, 13,* 493–509.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646–656.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2,* 1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects. *The Annals of Statistics, 6,* 34–58.

Rubin, D. B. (1986). Which ifs have causal answers. *Journal of the American Statistical Association, 81,* 961–962.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice, *Biometrics, 52,* 249–264.

Sackett, D. L., & Gent, M. (1979). Controversy in counting and attributing events in clinical trials. *New England Journal of Medicine, 301,* 1410–1412.

Schachter, S. (1959). *The psychology of affiliation.* Stanford, CA: Stanford University Press.

Schwarz, N. B., & Hippler, H. J. (1995). Subsequent questions may influence answers to preceding questions in mail surveys. *Public Opinion Quarterly, 59,* 93–97.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51,* 515–530.

Seaver, W. B., & Quarton, R. J. (1976). Regression discontinuity analysis of the dean's list effects. *Journal of Educational Psychology, 68,* 459–465.

Sechrest, L., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments. In L. Sechrest, S. G. West, M. A. Phillips, R. Rodner, & W. Yeaton (Eds), *Evaluation studies review annual* (Vol. 4, pp. 15–35). Beverly Hills, CA: Sage.

Shadish, W. R. (in press). The empirical program of quasi-experimentation. In L. Bickman (Ed.), *Contributions to research design: Donald Campbell's legacy* (Vol. 2). Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (in press). *Experimental and quasi-experimental design for generalized causal inference.* Boston: Houghton-Mifflin.

Shadish, W. R., Hu, X., Glaser, R. R., Knonacki, R., & Wong, S. (1998). A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychological Methods, 3,* 3–22.

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment to psychotherapy experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology, 64,* 1290–1305.

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

Sörbom, D. (1979). An alternative to the methodology for analysis of covariance. In K. G. Jöreskog & D. Sörbom (Eds.), *Advances in factor analysis and structural equation models* (pp. 219–234). Cambridge, MA: Abt Books.

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence–accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118,* 315–327.

"Student" (W. S. Gosset). (1931). The Lanarkshire milk experiment. *Biometrika, 23,* 398–406.

Stolzenberg, R. M., & Relles, D. A. (1990). Theory testing in a world of constrained research design. *Sociological Methods and Research, 18,* 395–415.

Swingle, P. G. (Ed.). (1973). *Social psychology in natural settings.* Chicago: Aldine.

Tebes, J. K., Snow, D. L., & Arthur, M. W. (1992). Panel attrition and external validity in the short-term follow-up study of adolescent substance use. *Evaluation Review, 16,* 151–170.

Trochim, W. M. K. (1984). *Research design for program evaluation: The regression-discontinuity approach.* Beverly Hills, CA: Sage.

Trochim, W. M. K., & Cappelleri, J. C. (1992). Cutoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials, 13,* 571–604.

Trochim, W. M. K., Cappelleri, J. C., & Reichardt, C. S. (1991). Random measurement error does not bias the treatment effect estimate in the regression-discontinuity

design: II. When an interaction effect is present. *Evaluation Review, 15,* 571–604.

Velicer, W. F., & Colby, S. M. (1997). Time series analysis of prevention and treatment research. In K. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 211–250), Washington, DC: American Psychological Association.

Velicer, W. F., & Harrop, J. W. (1983). The reliability and accuracy of time series model identification. *Evaluation Review, 7,* 551–560.

Velicer, W. F., & MacDonald, R. P. (1984). Time series analysis without model identification. *Multivariate Behavioral Research, 19,* 33–47.

Vinokur, A. D., Price, R. H., & Caplan, R. D. (1991). From field experiments to program implementation: Assessing the potential outcomes of an experimental intervention program for unemployed persons. *American Journal of Community Psychology, 19,* 543–562.

Virdin, L. M. (1993). *A test of the robustness of estimators that model selection in the nonequivalent control group design.* Unpublished doctoral dissertation, Arizona State University.

Weiner, B. (Ed.). (1974). *Achievement motivation and attribution theory.* Morristown, NJ: General Learning Press.

West, S. G., & Aiken, L. S. (1997). Towards understanding individual effects in multiple component prevention programs: Design and analysis strategies. In K. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 167–210). Washington, DC: American Psychological Association.

West, S. G., Aiken, L. S., & Todd, M. (1993). Probing the effects of individual components in multiple component prevention programs. *American Journal of Community Psychology, 21,* 571–605.

West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality, 59,* 611–662.

West, S. G., Hepworth, J. T., McCall, M. A., & Reich, J. W. (1989). An evaluation of Arizona's July 1992 drunk driving law: Effects on the city of Phoenix. *Journal of Applied Social Psychology, 19,* 1212–1237.

West, S. G., Newsom, J. T., & Fenaughty, A. M. (1992). Publication trends in *JPSP*: Stability and change in the topics, methods, and theories across two decades. *Personality and Social Psychology Bulletin, 18,* 473–484.

West, S. G., & Sagarin, B. (in press). Subject selection and loss in randomized experiments. In L. Bickman (Ed.), *Contributions to research design: Donald Campbell's legacy* (Vol. 2, pp. 117–154). Thousand Oaks, CA: Sage.

Wilkinson, L. and the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of change. *Psychological Bulletin, 116,* 363–381.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Wolchik, S. A., West, S. G., Sandler, I. N., Tein, J.-Y., Coatsworth, D., Lengua, L., Weiss, L., Anderson, E. R., Greene, S. M., & Griffin, W. A. (in press). An experimental evaluation of theory mother and mother-child programs for children of divorce. *Journal of Consulting and Clinical Psychology.*