

Participant Selection and Loss in Randomized Experiments

Stephen G. West

Brad J. Sagarin

Applied researchers taking a public health orientation often wish to answer questions about the effects of a treatment on a defined population of participants. To illustrate, they may wish to know whether a new substance use prevention program lowers the risk that adolescents will use drugs, what proportion of men who have had a myocardial infarction (MI) could avert a second one by taking aspirin, or what effect 30 minutes of daily aerobic exercise would have on a measure of quality of life in women 40-65 years of age. To answer such questions, statisticians (e.g., Draper, 1995; Kish, 1987) have recommended a formal two-stage statistical model, presented in Figure 6.1. In the first stage of this model, a random sample of participants is selected from a defined population. In the second stage, this sample of participants is randomly assigned to treatment and comparison (control) conditions. The purpose of the first stage (A) is to ensure that the results in the sample will represent the results in the population within a defined level of sampling error. Campbell and his associates (Campbell, 1957; Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish, Cook, & Campbell, in press) have termed this desideratum *external validity*. The purpose of the second stage (B) is to ensure that the observed effect on the dependent variable is due to some aspect of the treatment rather than other confounding factors such as maturation, history, or selection. Campbell and his associates have termed this desideratum *internal validity*. The beauty of this formal

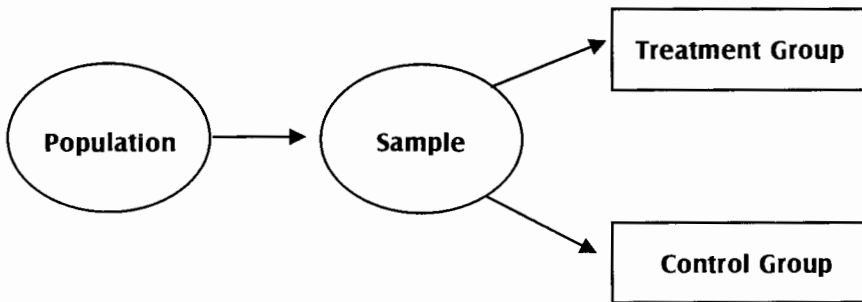
A. Random Sampling**B. Randomization**

Figure 6.1. 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.

two-stage model is that it ensures that internal and external validity will be simultaneously maximized.

Although a thing of beauty, the formal two-stage model has almost never been implemented in practice by applied (or basic) researchers in the behavioral sciences. Researchers in health and mental health frequently implement randomized experiments testing treatments of real policy interest; however, these randomized trials are nearly always implemented on a convenience sample of volunteers. Members of the sample often meet eligibility criteria ensuring that they are selected from the population of interest rather than some other population, but comparisons of the achieved sample with the population of interest are rare. Even rarer are studies that attempt to adjust estimates of treatment effects based on the characteristics of the achieved sample so that they can be generalized to the defined population. Furthermore, those individuals who do participate in a randomized experiment typically are not as compliant as some classical statistical presentations of experimental design would have us believe. Not all participants who are randomly assigned to exercise 30 minutes per day do so, for example, nor will all participants complete the dependent measure of quality of life. In a long-term investigation, the logistics of carrying out the planned measurements on all participants become difficult or even impossible: Some participants will have moved out of the geographic area, some will simply refuse, some will not be locatable, and some will have died. Because not all participants will receive the intended treatment and not all participants will complete the dependent measure, even the best-implemented randomized experiments will only approximate the ideal of the two-stage statistical model.

The purpose of this chapter is to address these issues from two perspectives. First, we draw on recent work on these problems in the statistics literature. This work builds on the twin foundations of missing data theory (e.g., Little & Rubin, 1987) and Rubin's (1974, 1978) approach to causal inference in experiments. We provide a nontechnical introduction to the basic ideas of a number of new design and analysis strategies that can be used in the face of problems of treatment noncompliance, participant attrition, and nonrandom sampling of participants. These approaches promise to strengthen the internal validity and external validity of inferences in the randomized experiment. Second, we draw on the work on internal and external validity by Campbell and his associates (Campbell, 1957, 1986; Campbell & Stanley, 1966; Cook, 1993; Cook & Campbell, 1979; Shadish et al., in press) that has been developed over the past four decades. This work has attempted to take a far more practical approach to these issues, collecting and systematizing insights from researchers to develop principles to guide the design and critical analysis of research. As we will see, these principles complement the statistical approaches and help shed light on several remaining issues in these areas.

Toward this end, we begin by reviewing ideas from work on missing data theory (Little & Rubin, 1987) and Rubin's approach to causality in experiments, which has come to be termed the Rubin causal model (Holland, 1986; Rubin, 1974, 1978). We first apply these central ideas to the problem of treatment noncompliance in randomized experiments in which the outcome variable for all responses is measured at a single posttest. We then extend these ideas to the problem of attrition in which the dependent variable cannot be measured at posttest on a subsample of participants. Finally, we address issues of selection into experiments and consider how better quantitative and qualitative estimates of treatment effects for the population of interest might be developed, drawing on Campbellian analyses of internal and external validity.

MISSING DATA MECHANISMS

Rubin's (1976) theoretical analysis of missing data emphasized consideration of the mechanism that is producing the missing data. For ease of exposition, we will differentiate between (a) independent variables (X s) that are measured at pretest (covariates), (b) manipulated treatment variables, and (c) the outcome variable of interest measured at posttest (Y). Rubin identified three classes of missing data mechanisms that have different implications for data analysis.

First, data may be *missing completely at random* (MCAR). In this case, the probability that data on the outcome variable are missing is independent of any manipulated treatment variable or measured covariate, as well as being independent of the value of the dependent variable. For example, consider the earlier example of a substance abuse prevention program for adolescents. To form one

outcome variable, a saliva sample was collected for pharmaceutical analysis from each participant at the completion of the experiment. Suppose now that a clumsy laboratory assistant dropped a box containing 10% of the samples, breaking all of them. If the collection tubes had been arranged in a random order, the missing (lost) samples would be expected to be independent of any characteristic of the participants. In this example, the source of the missing data is a random event that is unrelated to all other independent variables. Such cases do not create any special problems for analysis, generalization, or causal inference.¹ Estimates of treatment effects will be unbiased, as will generalization of results based on the sample of available cases to the full sample. No adjustment of the results is necessary.

Second, data may be *missing at random* (MAR). In this case, the probability that data are missing depends only on one or more X variables that are measured on all participants in the data set. If these variables are related to the outcome variable, then the data are described as missing at random. To illustrate, consider our substance abuse prevention experiment. Suppose that mathematically gifted students are assigned to a special mathematics enrichment program based only on their achievement test scores. Suppose further that the scheduling of enrichment program classes precludes participation of these students in the treatment of interest, the substance abuse prevention program. If we assume that all other students receive the program, then the only source of missingness is the student's score on the mathematics achievement test. Data that are missing at random potentially lead to biased means and variances. Such data potentially require adjustment for the values of the X variables that are related to missingness to produce unbiased estimates of treatment effects, generalization of treatment effects to the population, or both.

Third, there may be a *non-ignorable missing data mechanism*. In this case, the probability that the data are missing is related to the participant's (in some cases unmeasured) response on the outcome variable Y . To illustrate, participants who are heavy substance users may be more likely to miss data collection sessions in a drug abuse prevention experiment. Data that would be collected potentially lead to biased estimates of all parameters that may be of interest—means, variances, and unstandardized regression coefficients. In the absence of knowledge of the missing data mechanism, any attempt to adjust estimates of treatment effects or to generalize treatment effects to the population will be fraught with uncertainty.

A variety of useful statistical techniques have been developed over the past two decades for addressing problems of missing data (see Arbuckle, 1996; Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997; Little & Rubin, 1990; and Little & Schenker, 1995, for introductions and Little & Rubin, 1987; Rubin, 1987; and Schafer, 1997, for more advanced treatments). These techniques include the EM algorithm, multiple imputation, and maxi-

imum likelihood approaches. These strategies produce far better estimates in many applications than earlier traditional strategies such as listwise deletion, pairwise deletion, or mean substitution (see Cohen & Cohen, 1983). When data are MCAR or MAR, these new statistical techniques provide unbiased estimates of all population parameters of interest. Nonetheless, researchers can rarely be sure that their data are MCAR or MAR. Consequently, as Cook and Campbell (1979) suggest, perhaps the best strategy for addressing missing data is to incorporate features at the design phase that either help minimize their occurrence or that help provide an understanding of the mechanisms that produce missing data in the specific research context (i.e., that turn non-ignorable missing data into MAR data).

THE RUBIN CAUSAL MODEL: AN INTRODUCTION

The Rubin causal model (Angrist, Imbens, & Rubin, 1996; Holland, 1986; Rubin, 1974, 1978) begins with consideration of the ideal conditions that would be required to observe a causal effect. These ideal conditions would permit comparison of the participant's response under the treatment condition with what *would* have happened under the comparison condition under identical circumstances. Under these hypothetical ideal conditions, the causal effect of the treatment could be precisely defined as $Y_t(u) - Y_c(u)$, where Y is the observed response, t refers to the treatment condition, c refers to the comparison condition, and u is the unit (typically an individual participant). As noted by Holland (1986), however, these hypothetical ideal conditions cannot precisely occur in practice, leading to the fundamental problem of causal inference: "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 " (p. 947). Researchers may, however, *infer* causality if they are willing to make specific assumptions. In field experiments, two assumptions are necessary for the inference of causality. First, assignment of participants to treatment conditions is assumed to be independent of all other variables. Researchers attempt to meet this assumption through randomization of participants to treatment and comparison conditions. Successful randomization guarantees that, on average, the participants in the treatment group and the participants in the control group can be expected to be equal on all possible background variables at pretest. More formally, $E(\bar{X}_t) = E(\bar{X}_c)$ for any variable X (including the pretest measure of the outcome variable Y) in the absence of treatment. Second, researchers must assume that (a) the process of randomization has no impact on the participant's responses and (b) the participant's responses are not affected by the treatments received by other participants. Rubin has collectively labeled these conditions as the Stable Unit Treatment Value Assumption (SUTVA). The first part of this assumption is typically

considered to be plausible in randomized experiments; the second part is normally addressed by research procedures that minimize contact among participants, particularly among those in different treatment groups (see Cook & Campbell, 1979). Given that (a) treatment assignment is independent of all background variables and (b) SUTVA holds, $\bar{Y}_t - \bar{Y}_c$ will provide an unbiased estimate of the average causal effect of the treatment in the population of interest. Because Rubin demands very precise specification of all aspects of the experiment, however, the generalization of this causal inference will be very limited. The causal inference is limited to the *specific* treatment and comparison conditions, the *specific* population from which the random sample is drawn, and the *specific* dependent measure collected in the experiment. Rubin's model provides few guidelines for generalizing the estimate of the treatment effect to other treatment conditions and outcomes beyond the specific instances that are studied, a point that will become important later in this chapter.

Some additional insight into the randomized experiment can be gained from reconsidering the randomized experiment from the perspective of missing data theory. Figure 6.2A provides the starting point for this consideration, illustrating Rubin's *ideal* situation for observing causality. In this depiction, we are able to observe each participant's response to the treatment and control conditions under identical circumstances. In this case, we see that the true causal effect is $Y_t(u) - Y_c(u) = 1$ for each of the participants, representing a constant effect of treatment. In Figure 6.2B we consider a more realistic situation of data that would actually be available from a randomized experiment. The second column indicates the random assignment of each participant to either the treatment or comparison group: six participants are assigned to the treatment group (t) and six to the comparison group (c). Each participant's response is indicated only for the treatment condition to which he or she is assigned; the participant's response under the other condition is not observed (missing data). Given random assignment to treatment and comparison conditions, treatment assignment is expected to be independent of all participant characteristics so that the unobserved data are missing completely at random. Consequently, the means and variances in each treatment group will be unbiased estimates of the corresponding population parameters, so that $\bar{Y}_t - \bar{Y}_c$ will be an unbiased estimate of the true treatment effect in the population. No adjustment of this treatment effect is necessary. Any discrepancy between the true causal effect and the estimated causal effect in a single experiment is the result of random error in the assignment of participants to treatment conditions. In any actual sample, $\bar{Y}_t - \bar{Y}_c$ typically will not be exactly 0 in the absence of a treatment effect. Consequently, the sample-based estimate will fluctuate probabilistically around the true value of the effect size as different samples are selected.

A. Rubin's Ideal Model

<i>Participant</i>	Y_t	Y_c
1	3	2
2	4	3
3	5	4
4	5	4
5	6	5
6	7	6
7	3	2
8	4	3
9	5	4
10	5	4
11	6	5
12	7	6
	$\bar{Y}_t = 5.0$	$\bar{Y}_c = 4.0$

B. Randomed Experiment

<i>Participant</i>	A	Y_t	Y_c
1	0	■	2
2	0	■	3
3	1	5	■
4	0	■	4
5	1	6	■
6	0	■	6
7	1	3	■
8	0	■	3
9	1	5	■
10	1	5	■
11	0	■	5
12	1	7	■
		$\bar{Y}_t = 5.16$	$\bar{Y}_c = 3.83$

Figure 6.2. Illustrative Applications of Rubin's Causal Model

NOTE: Entries in the column labeled A represent the treatment condition to which the participant is assigned (1=treatment; 0=comparison). Y_t is the response observed in the treatment condition; Y_c is the response observed in the comparison condition. The black squares indicate that the response was not observed.

TREATMENT NONCOMPLIANCE

To this point, we have assumed that all participants actually received the treatment to which they were assigned. In many experiments, however, a portion of the participants fail to follow the treatment protocol, a problem termed *treatment noncompliance*. For example, some participants who are assigned to complete 30 minutes of aerobic exercise per day may drop out of treatment, in effect reassigning themselves to the no-exercise control condition. The central issue is whether the typically unknown mechanism producing this treatment reassignment is ignorable.

Meier (1991) described a number of non-ignorable ways in which treatment compliance may relate to outcomes. Both outcomes and compliance may be affected by the same dispositional traits. For example, in our example of a substance abuse prevention program, highly rebellious adolescents may have a greater predisposition toward drug use and be more likely to skip the prevention program. Alternatively, characteristics of the treatment may lead to noncompliance. For example, a particularly coercive or demanding treatment may lead to greater levels of noncompliance than an innocuous control condition. To complicate the matter still further, outcomes themselves may cause changes in compliance. For example, those adolescents in the drug prevention program condition who begin using drugs may skip the program to avoid reprimand by program personnel. Punctuating the difficulty, Meier concludes, "Worse yet, these phenomena may interact in complex ways, and it may be that noncompliance is dominated by different mechanism in different groups" (p. 21). Given these issues, any statistical approach, if it is to be successful, must address the possibility that treatment compliance is non-ignorable.

STATISTICAL APPROACHES TO NONCOMPLIANCE

We begin this section with the simplest case of noncompliance, in which participants either receive the full treatment condition or the full no treatment (control) condition. We then consider approaches in which each participant's degree of compliance is measured as a continuous variable. Each of the approaches considered in this section assumes that all participants are measured on the outcome variable at posttest.

Dichotomous Measures of Treatment Compliance

When treatment compliance is represented as a dichotomous variable, three general classes of statistical approaches have been taken to noncompliance:

intention to treat analysis, analysis by treatment received, and a variety of new approaches that directly attempt to address the possibility that compliance is non-ignorable. These three approaches reflect a tension between the maintenance of group comparability and legitimate causal inference, on one hand, and the estimation of the effect of the actual treatment on the other.

Intention to Treat Analysis

In intention to treat (ITT) analysis, the issue of treatment noncompliance is ignored. Instead, Fisher's maxim of "analyze them as you've randomized them" (Fisher, cited in Boruch, 1997, p. 195) is followed. ITT's major advantage is its maintenance of the integrity of random assignment, which permits causal inferences to be made without making assumptions beyond those required for the randomized experiment (Lee, Ellenberg, Hirtz, & Nelson, 1991). Its major disadvantage is that it sacrifices the ability to estimate the effect of the treatment actually received. This disadvantage can be seen through an extension of our earlier illustration of Rubin's causal model.

Figure 6.3A adds treatment noncompliance to the earlier illustration. It begins with the data illustrated in Figure 6.2B but makes one important change for ease of presentation: The first half of the participants have been assigned to the treatment condition and the second half of the participants to a no treatment control condition. This change results in a matched case assignment that perfectly equates the two groups—participant 1 is identical to participant 7, participant 2 is identical to participant 8, and so on. By equating participants in this manner, we do not need to address random error in the assignment of participants to treatment conditions, thus simplifying the illustration. The true value of the estimate in this sample of the causal effect if Rubin's ideal case could be implemented is 1.0. The third column in Figure 6.3A indicates whether participants complied with or failed to comply with their treatment assignment. Note that the first two participants (who are assigned to the treatment group) do not comply and receive no treatment; hence, we observe the same values of the outcome variable as we would have *if* they had been assigned to the no treatment control condition.

Applying the intention to treat analysis to the data in Figure 6.3, the mean of the observed responses for the first six participants (assigned to the treatment group) is compared with the mean of the observed responses for the second six participants (assigned to the control group). The difference between these two means, here 0.67, is the ITT estimate of the treatment effect. Note that, following the logic of the Rubin causal model in this case, the true effect of treatment when it is actually delivered is 1.0. This discrepancy between the two estimates of the treatment effect is a systematic effect of treatment noncompliance. Recall

that we set up our example so that participants 1-6 are perfectly matched with participants 7-12, thus eliminating any possibility of random error in the assignment of participants to treatment conditions.

According to critics of ITT analysis (e.g., Goetghebeur & Shapiro, 1996; Sheiner & Rubin, 1995; Sommer & Zeger, 1991), the estimate of the treatment effect provided by ITT analysis may prove to be of little value. Sheiner and Rubin argue that the ITT estimate of the treatment effect confounds both the efficacy of the treatment and the effectiveness of the instructions to comply with the treatment in the context of the specific experimental trial. This confounding means that ITT is likely to provide a biased estimate of the treatment effect. Typically, this bias will lead to a conservative estimate of the treatment effect, as in our illustration. ITT can also lead to overestimation of the average treatment effect if the treatment effect is not constant for all participants, with some participants having less positive outcomes in the treatment condition than they would have in the control condition.

Figure 6.3B provides an illustration of this latter case. Notice that the treatment has a negative effect for participants who have a low value of Y in the absence of treatment and a positive effect for participants who have a high value of Y in the absence of treatment. This case could occur, for example, if there were a Baseline (pretest) Level of $Y \times$ Treatment interaction effect with a cross-over form. According to the Rubin causal model, the estimate of the average causal effect in Figure 6.3B is 0, yet the intention to treat analysis overestimates the causal effect, producing an estimate of 0.5. Thus, although the intention to treat analysis is very likely to produce an underestimate of the true treatment effect, there is no guarantee that it will do so in any particular experiment.

Some advocates have argued that the ITT analysis does produce a useful estimate of the treatment effect that may be generalized to the population. They argue that treatment efficacy and treatment compliance are inevitably confounded when the intervention is implemented on a large-scale basis. Consequently, ITT estimates of treatment effects may include a realistic reflection of the degree of noncompliance that normally manifests itself outside the setting of the randomized trial. Of course, the viability of such arguments strongly depends on the degree to which the amount of noncompliance observed in the randomized trial closely mirrors that observed when the treatment is actually implemented in the population. Available research on medical compliance and adherence (e.g., Gochman, 1997; Kaplan & Simon, 1990; Meichenbaum & Turk, 1987) suggests that, if anything, (a) individuals who are selected and who agree to participate in randomized trials may be more motivated to comply than the typical member of the population, and (b) these individuals may further increase their level of compliance as a result of the high levels of monitoring of their participation and outcomes and support for the regimen by the researchers during the trial. In contrast, other researchers have argued that once a treatment

A. Constant Treatment Effect

<i>Participant</i>	<i>A</i>	<i>C</i>	Y_t	Y_c
1	1	0	3	2*
2	1	0	4	3*
3	1	1	5*	4
4	1	1	5*	4
5	1	1	6*	5
6	1	1	7*	6
7	0	1	3	2*
8	0	1	4	3*
9	0	1	5	4*
10	0	1	4	4*
11	0	1	6	5*
12	0	1	7	6*

B. Non-Constant Treatment Effect

<i>Participant</i>	<i>A</i>	<i>C</i>	Y_t	Y_c
1	1	0	0	2*
2	1	0	2	3*
3	1	1	4*	4
4	1	1	4*	4
5	1	1	6*	5
6	1	1	8*	6
7	0	1	0	2*
8	0	1	2	3*
9	0	1	4	4*
10	0	1	4	4*
11	0	1	6	5*
12	0	1	8	6*

Figure 6.3. Illustration of Treatment Noncompliance

NOTE: Entries in the column labeled A represent the treatment condition to which the participant is assigned (1 = treatment; 0 = comparison). Entries in the column labeled C indicate whether the participant complied with the treatment assignment (1 = yes; 0 = no). Y_t is the response observed in the treatment condition; Y_c is the response observed in the comparison condition. Asterisks indicate the response that was actually observed.

is proven, participants may be more willing to comply (Robins & Greenland, 1996), or an alternative treatment delivery method may ensure 100% compliance (Sommer & Zeger, 1991). The resolution of these arguments depends, in part, on an analysis of the generalization of the observed compliance effect (as well as the treatment effect) from the treatments, settings, and participants used in the randomized trial to the treatments, settings, and participants that constitute the target populations for generalization (Cook, 1993; Cronbach, 1982), a topic to which we will return later in this chapter.

Analysis by Treatment Received

In the analysis by treatment received (TR), the analyst ignores the original assignment of participants to treatment and control groups. Instead, the data are analyzed according to the treatment the participant actually received.² Although the TR approach permits a numerical comparison between participants who received the treatment and those who did not, the participants have now self-selected into treatment and control groups, so that these groups can no longer be expected to be comparable in the absence of a treatment effect. Consequently, analysis by treatment received is presumed to produce biased estimates of treatment effects, with the direction and magnitude of bias being unpredictable (see Lee et al., 1991; Sheiner & Rubin, 1995). To illustrate the TR approach, participants 1, 2, and 7-12 in Figure 6.3A actually receive the control treatment, with $\bar{Y}_c = 3.625$ for these eight participants. Participants 3-6 actually receive the treatment program, with $\bar{Y}_t = 5.75$ for these four participants. The estimate of the causal effect, $\bar{Y}_t - \bar{Y}_c = 2.125$, provides a substantial overestimate of the true value of 1.0 obtained from the Rubin causal model in this case.

New Statistical Approaches

The ITT approach only considers information about treatment assignment; the TR approach only considers information about treatment compliance. In contrast, a variety of statistical approaches have recently been developed that consider information from both sources. In the context under consideration in which compliance is considered to be a dichotomous variable (comply, noncomply), one of two different types of estimates of treatment effects may be produced. One set of approaches (e.g., Angrist et al., 1996) makes a number of critical assumptions and yields an estimate of the treatment effect for the compliers, termed the *complier average causal effect* (CACE). Estimates of CACE are unbiased assuming the assumptions are adequately met; they can be computed for either dichotomous or continuous outcome variables. A second set of approaches (e.g., Robins, 1989) attempts to consider the degree of uncertainty in the estimation of the treatment effect, resulting in estimates of the range of

possible effect sizes for compliers and noncompliers. These estimates may be computed only for outcome variables that are dichotomous or that represent the time elapsed before a dichotomous outcome occurs (failure time or survival models).

The Complier Average Causal Effect. Angrist et al. (1996) present one of the most complete developments of the central ideas of the CACE estimate. They begin by categorizing participants into four groups based on their compliance patterns. *Compliers* are participants who would comply with treatment if assigned to the treatment group and who would not seek out treatment if assigned to the control group. In our exercise example, these are women who would perform 30 minutes of aerobic exercise if assigned to the treatment condition and no exercise if assigned to the control condition. *Never-takers* are participants who would refuse treatment regardless of treatment assignment. In our exercise example, these are women who would not exercise regardless of their treatment assignment. *Always-takers* are participants who would seek out treatment regardless of treatment assignment. In our exercise example, these are women who would always do 30 minutes of aerobic exercise regardless of their treatment assignment. Finally, *defiers* are participants who would refuse treatment if assigned to the treatment group but who would seek out treatment if assigned to the control group. In our exercise example, defiers would refuse to exercise if assigned to the treatment condition and exercise 30 minutes per day if assigned to the control condition.

Several strong statistical assumptions must be made for the CACE estimate to be unbiased. Angrist et al. (1996) and Imbens and Rubin (1997) start with the two assumptions discussed previously that are required for the estimation of a causal effect in any experiment: randomization and SUTVA. They then note that two additional assumptions must be satisfied as well.

The first is the *weak exclusion restriction* (Imbens & Rubin, 1997), which requires that treatment assignment have no effect on outcomes for never-takers and always-takers. Otherwise stated, those participants whose treatment is not affected by assignment should not have their outcomes affected by assignment either. This assumption implies that the experience of control group participants must be identical to the experience of noncompliant participants in the treatment group. For example, consider the studies of the effect of Vietnam-era veteran status on health outcomes, which approximate randomized experiments because young men's draft status was determined through a draft lottery mechanism (Angrist et al., 1996). The weak exclusion restriction would require identical health outcomes for (a) always-takers who were drafted and always-takers who enlisted voluntarily and (b) never-takers who were not drafted and never-takers who dodged the draft. Condition (a) seems plausible, but given the potentially life-changing effects of dodging the draft (e.g., migra-

tion to another country, prison terms, health-impairing actions), condition (b) is less so. Angrist et al. (1996) note that the lower the correlation between treatment assignment and treatment received, the greater the proportion of noncompliers and the greater the bias stemming from violations of the weak exclusion restriction.

The second assumption is that of *monotonicity*. The monotonicity assumption requires that there be no defiers—no participants who would specifically obtain treatment if assigned to the control group and refuse treatment if assigned to the treatment group. This assumption permits the unambiguous classification of noncompliers in the treatment and control groups. Noncompliers assigned to the treatment group can then be classified as never-takers, whereas noncompliers in the control group can be classified as always-takers. Violation of the monotonicity assumption leads to greater bias as the proportion of defiers increases. Similarly, the amount of bias will be greater as the difference between the effect of treatment on compliers and defiers increases³ (Angrist et al., 1996).

Given these assumptions, the CACE estimate—the unbiased estimate of the effect of treatment in the specific subpopulation of compliers—may be produced. To provide a conceptual illustration of the CACE approach, let us reconsider our simple example in Figure 6.3A in which participants 1 and 2 are never-takers and there are no always-takers or defiers. To compute a treatment effect for compliers, we clearly need to discard participants 1 and 2—these are never-takers who did not comply with their treatment assignment. Recall that participants 7 and 8 are matched to participants 1 and 2, so we also need to discard these never-taker participants who would not have complied *if* they had been assigned to the treatment group. Discarding the never-taker participants in both the treatment and control groups provides an unbiased estimate of the treatment effect, here 1.0, for the subgroup of participants who are compliers—participants who would follow the protocol regardless of whether they were assigned to the treatment or control condition. Note also what happens when we apply this same approach to the example of a nonconstant treatment effect in Figure 6.3B. The CACE estimate is the difference between the mean of the scores of the four compliers in the treatment group ($\bar{Y}_t = 5.50$) and the mean of the four compliers in the control group ($\bar{Y}_c = 4.75$). This value of 0.75 corresponds to the average treatment effect from the Rubin causal model for the four participants in each group who are compliers. It does not correspond to the estimate of an average treatment effect of 0.0 from the Rubin causal model when the full set of six participants in the treatment group ($\bar{Y}_t = 4.0$) and six participants in the control group ($\bar{Y}_c = 4.0$) is considered. This example helps illustrate that the CACE estimate provides an unbiased estimate of the average treatment effect only for the population of compliers; CACE estimates do not necessarily generalize to the full population.

The central challenge for the CACE approach is that the compliance category status of participants in the control group is unknown—we normally cannot identify which control group participants would comply with the treatment protocol *if* they had been assigned to the treatment group. On the other hand, we know the overall proportion of the participants in the treatment group who complied with the treatment protocol. Thus, we can conceive of the known intention to treat estimate of the treatment effect as representing a mixture of two groups: (a) a group of compliers who would show a treatment effect of unknown size and (b) a group of never-takers who would show a treatment effect of 0. One simple method of providing an estimate of the treatment effect for the subgroups of compliers was suggested by Bloom (1984). Bloom's estimate takes the intention to treat estimate of the treatment effect, 0.67 in Figure 6.3A, and multiplies it by the inverse of the proportion of compliers, here $(4/6)^{-1}$. In the example presented in Figure 6.3A, this value is 1.0, exactly equal to the causal effect for compliers expected from the Rubin causal model (see Figure 6.2A).

Although Bloom's (1984) work provided an initial approach to the problem of noncompliance with continuous variables (see also Sommer & Zeger, 1991, for an analogous approach with dichotomous outcome variables), it has been largely superseded by other approaches that are applicable to both continuous and dichotomous variables. Angrist et al. (1996) extended Bloom's approach by more precisely articulating the necessary assumptions underlying CACE estimates and extending the model so that researchers may consider contexts in which both never-takers and always-takers may be present. Little and Yau (1998) developed a maximum likelihood-based estimation procedure that can yield improved large-sample estimates of CACE in contexts in which there are a large number of noncompliers. Imbens and Rubin (1997) developed a Bayesian estimate of CACE that yields improved estimates in small samples and permits exploration of how the results would change if the weak exclusion restriction and monotonicity assumptions are violated. Computer programs have been developed that implement these procedures; some are now publicly available (e.g., see Little & Yau, 1998).

In summary, the CACE approach provides an unbiased estimate of the treatment effect for participants classified as compliers, an estimate that may be preferred in many research contexts to the ITT or particularly the TR estimates (Goetghebuer & Shapiro, 1996; Sheiner & Rubin, 1995). Angrist et al. (1996) argue that the CACE estimate is particularly useful because "the average over the subpopulation of those whose behavior can be modified by assignment are [*sic*] likely to be informative about population averages of those who comply in the future" (p. 450). Nonetheless, the CACE approach is not without its limitations. The approach is based on a series of strong assumptions: successful randomization of participants to treatment and control conditions, SUTVA, the weak exclusion restriction, and monotonicity. When these assumptions are vio-

lated, the estimates of treatment effects can be severely biased. Although Imbens and Rubin (1997) propose a Bayesian method that can relax the weak exclusion and monotonicity assumptions, the resulting estimates of treatment effects are often very imprecise (associated with very large confidence intervals) and may be of little practical value. The CACE approach presumes, in effect, that some participants reassign themselves to the complete version of the other treatment group: Never-takers receive the control treatment and always-takers receive the full version of the active treatment. Thus, CACE may not yield proper estimates in situations in which two active treatments are compared (Robins & Greenland, 1996) or in which always-takers in the control condition get a lower dose of the active treatment (e.g., 15 minutes of exercise per day) than always-takers in the treatment condition. CACE presumes that compliance is truly dichotomous: When compliance is continuous, as when participants in the treatment condition attend between 0% and 100% of the sessions, CACE may yield different estimates depending on the threshold (e.g., number of sessions) chosen to define compliance. In some cases (e.g., Vinokur, Price, & Caplan, 1991), sensitivity analyses may show that the choice of this threshold makes little difference; in other cases, approaches that explicitly model the degree of compliance may be needed (see next section). In addition, the CACE approach presently assumes that the outcome variable is measured on all participants at posttest. Most experiments that extend over time involve attrition, sometimes of a substantial degree. Although Little and Yau (1998) have begun to address this issue, their approach to noncompliance with missing data on the outcome variable is still in development. Finally, generalization of the CACE estimate of the treatment effect makes the strong assumption that participants who comply with the treatment protocol in the context of the experimental setting will be identical to those who comply in the context of the applied settings that are the target of generalization. In cases in which there is a nonconstant treatment effect (e.g., see Figure 6.3B), changes in the participants who are compliers versus never-takers can potentially lead to dramatic changes in the estimate of the treatment effect.

Circumscribing the Effect of Treatment on Everyone. A second approach for addressing noncompliance estimates what Robins and Greenland (1996) have termed the *global average treatment effect* (ATE). This approach estimates the bounds on (i.e., theoretical range of) possible treatment effect sizes for all participants based on both the observed outcome data and assumptions about the data that are not observed. This approach is applicable in randomized experiments with dichotomous levels of treatment assignment, a dichotomous measure of treatment received, and either a dichotomous or time-to-failure measure of outcomes (Robins, 1989).

ATE produces bounds for an estimate of great inherent interest to most researchers: the effect of treatment on the population of *all* participants, not just compliers. One of the seeming disadvantages of the ATE estimate is that the range of possible treatment effects may be quite wide. Robins and Greenland (1996), argue, however, that "wide bounds make clear that [*sic*] the degree to which public health decisions are dependent on merging the data with strong prior beliefs" (p. 457). On the positive side, they further note that, at times, the lower bound can lie well above zero, indicating a clear treatment effect. In addition, Balke and Pearl (1994) offer a technique to determine the specific distribution of the outcome in the population that leads to a particular value within the bounds. Some of these distributions may be quite unrealistic, so that these values can be eliminated (e.g., in our earlier myocardial infarction example, a distribution in which *all* noncomplying participants in the treatment condition who fail to take aspirin would die within 5 years if they had complied with treatment). Thus, an examination of values that produce implausible distributions can enable the researcher to substantially tighten the bounds on the treatment effect estimate, changing the bounds from those that are mathematically possible to those that are realistic. Balke and Pearl (1997) provide sharp ATE bounds that, in certain circumstances, improve substantially over the bounds presented in Robins (1989).

In contrast to the previously presented CACE approaches, the ATE approach does not theoretically require that researchers meet the assumptions of the weak exclusion restriction, monotonicity, or even random assignment. Once again, versions of the ATE approach that omit these assumptions carry a potential cost in terms of increased bounds on the ATE estimate. Robins (1989; Robins & Greenland, 1996) calculated ATE bounds for all combinations of these assumptions and found that lack of monotonicity did not appear to be critical; violations of this assumption had little, if any, effect on the bounds. In contrast, violation of the weak exclusion restriction and random assignment often led to substantially increased bounds. Balke and Pearl (1997) have also developed procedures for investigating the opposite question: Given an obtained pattern of data, can researchers infer which, if any, assumptions must have been violated. Balke and Pearl note that certain specified patterns of data point to violation of either the successful random assignment or the weak exclusion restriction assumption.

Continuous Measures of Compliance

In some contexts, researchers may collect a continuous measure of the degree of compliance with the treatment protocol. For example, in an educational intervention, researchers may record the proportion of the intervention sessions each participant attends or the proportion of the workbook materials each participant

completes. Drug studies may use counts of unused pills or technologies such as “smart” pill bottles that record each time the bottle is opened that develop a measure of compliance. Given that a high-quality continuous measure of compliance is available,⁴ statistical approaches are available that provide estimates of treatment effects conditioned on the participant’s level of compliance. Once again, strong assumptions must be met for the procedure to produce unbiased estimates of treatment effects.

One approach to the problem of continuous levels of compliance was proposed by Holland (1988). He addressed what has been termed the *encouragement design* (Powers & Swinton, 1984; see also Brewer’s, 1976, presentation of the randomized invitation design), in which continuous levels of compliance are observed in both the treatment and control groups. An encouragement design is a randomized experiment in which participants in the treatment group are encouraged to perform some activity, such as studying for an upcoming test or exercising on a regular basis. In the studying example, the researcher would record the number of hours each participant in both the treatment and control groups studied as well as his or her score on the outcome measure, here test performance. Holland developed the additive linearly constant effects (ALICE) model, which provides an estimate of the causal effect of each additional hour of study on test performance. This model begins by making the standard assumptions required for causal interpretation of treatment effects in an experiment, successful randomization and SUTVA. The ALICE model also requires that several additional assumptions are met: The effects of treatment assignment and treatment received on the outcome are additive, the effect of treatment received on the outcome is linear, and the effects are constant across participants (e.g., no interactions with participant characteristics). Finally, in order to provide a unique, unbiased estimate of the causal effect of treatment on the outcome, the analyst must make one of two further assumptions: (a) All students would score identically if they were not encouraged to study and did not study (i.e., 0 hours), regardless of how much they would actually study if not encouraged; or (b) “encouragement, of and by itself, has no effect on [the outcome]” (Holland, 1988, p. 470). The first of these assumptions typically will be plausible only if the outcome variable represents an entirely new knowledge base (or behavioral activity) on which all participants would be expected to receive a score of no prior knowledge (0). The second assumption is much more likely to be plausible in practice; however, it can be compromised to the extent that encouragement creates expectations of positive performance on the test (e.g., Rosenthal & Rubin, 1978).

When the second, more plausible assumption is made, the ALICE model produces a treatment effect estimate that is conceptually very similar to the CACE estimate discussed previously—the ALICE estimate is a ratio of the total effect of assignment on the outcome to the total effect of assignment on compliance.

Note that whereas the ALICE estimate assumes linearity, the CACE estimate is linear by definition because compliance may take on only two possible values. Note also that the (strong) exclusion restriction⁵ discussed by Angrist et. al. closely parallels Holland's assumptions of no direct effect of encouragement and additive (noninteractive) effects. The central distinction between the models is that Holland makes the assumption of constant effects, an assumption that allows the ALICE model to represent all participants. Angrist et al., on the other hand, do not make this assumption, so that their estimate of the causal effect will represent the *average* effect for the population of *compliers*—no statement is made about the unmeasured effect of treatment on *noncompliers*.

A second approach to continuous levels of compliance was developed by Efron and Feldman (1991) in the context of highly controlled, double-blind pharmaceutical experiments. In this context, participants in both the active drug and placebo conditions often display varying proportions of compliance with the prescribed treatment regimen. Efron and Feldman (1991) have developed a technique for estimating the dose-response curve, which describes the effect of varying levels of treatment dose on the outcome measure. Further, their approach relaxes Holland's assumption of linear effects, allowing for the modeling of nonlinear dose-response curves. Efron and Feldman present a complex empirical example in which the treatment group displayed a nonlinear dose-response relationship.⁶ The control group displayed a linear relationship between degree of compliance and outcomes. Given that the control group ingested pharmaceutically inactive placebos, the existence of a dose-response relationship in this group indicates that degree of compliance itself was related to outcomes.

The observed dose-response curve conceptually is obtained by subtracting the control group curve from the curve in the treatment group, providing an estimate of the actual dose-response curve. In the case of a *successful* double-blind experiment in which participants have no basis for inferring which drug they were taking, this estimate would provide an unbiased estimate of the causal effect at each level of compliance. Under these conditions, compliance level is assumed to be independent of the treatment condition. As Meier (1991) points out, however, there is no guarantee that the double blinding will be successful. Compliance may stem from different processes in each group (i.e., in the control group healthier patients discontinue taking placebos because they detect no benefit, whereas in the treatment group, sicker patients fail to comply because of intolerable side effects), so that treatment group assignment can no longer be expected to be independent of compliance level.

To estimate these models, researchers typically assume there is no interaction between the participant's pretest level on the outcome variable and level of compliance in determining the response.⁷ In the presence of such an interaction, Efron and Feldman's technique will produce an ambiguous range of possible

dose-response curves. In this case, the researcher can still either (a) estimate the actual dose-response curve by assuming the dose-response curve is linear within each condition, or (b) tighten the range of possible curves by utilizing variance information.

In applying this technique to data from a pharmaceutical trial, Efron and Feldman's analysis was complicated because the data displayed different distributions of compliance in the treatment and control groups. This observation cast doubt on the comparability of participants with similar levels of observed compliance. Efron and Feldman attempted to address this issue by adjusting the distribution of compliance in the control group to match the treatment group distribution. Thus, a participant at the 85th percentile of compliance in the control group was assigned an adjusted compliance score equal to the 85th percentile in the treatment group. As explicated by Meier (1991), this is justified if we regard observed compliance in each group as stemming primarily from a monotonic function of an "underlying propensity to be compliant" (p. 21). Albert and Demets (1994) explored the ramifications of this noncomparability in a simulation study, finding that "even moderate non-comparability . . . may produce severely biased estimates" (p. 2323).

Goetghebeur and Shapiro (1996) note that a researcher may wish to ask either of two questions: "What is the treatment benefit experienced by patients who were observed to have compliance level Z?" (i.e., self-selected levels of Z) or "How would the group as a whole have fared at treatment exposure level Z?" (p. 2814). Efron and Feldman's (1991) technique addresses the first question in all cases and also addresses the second question in the case of successful double blinding. Goetghebeur and Shapiro expand on the Efron and Feldman method in an effort to answer the second question under a wider range of conditions. Goetghebeur and Shapiro suggest methods of modeling side effects and propose methodological and statistical approaches for equating compliance in the treatment and control groups. Goetghebeur and Molenberghs (1996) suggest methods of estimating causal effects that are applicable when compliance is measured as (or categorized into) an ordered categorical variable and the outcome variable is dichotomous or is a continuous variable that has been dichotomized.

ATTRITION: PARTICIPANT LOSS AT POSTTEST MEASUREMENT

Participant attrition is a second problem that can lead to problems in generalizing estimates of the treatment effect to the population of interest (external validity) and even to biased estimates of causal effects in randomized experiments (internal validity). Attrition can and should be minimized by developing systems during the planning of the randomized trial for tracking the participants, maintaining participant interest and motivation throughout the experiment, and

providing participants with incentives for continued participation (Cook & Campbell, 1979; Ribisl et al., 1996). Nonetheless, attrition, sometimes of substantial magnitude, can be expected to occur even in the best-planned experiments. As one illustration, Biglan et al. (1991) reviewed longitudinal studies of substance abuse prevention programs and found that attrition rates ranged from 5% to 66% (mean = approximately 25%). Of concern, dropouts in these studies typically report greater substance abuse at the initial measurement, strongly implying that estimates of treatment effects will be biased if they are not corrected for the effects of attrition (McGuigan, Ellickson, Hays, & Bell, 1997).

Cook and Campbell (1979) note that attrition may take either (or both) of two forms that have distinct implications for the estimate of the causal effect of treatment. First, overall attrition may occur in which attrition is independent of the participants' assignment to the treatment or control groups. This form of attrition yields unbiased estimates of the average causal effect for the population who would complete the full experimental protocol, including posttest measurement. Generalization of this treatment effect may be limited, as link A in Figure 6.1 may no longer reflect a random sample of the full population of interest. Second, differential attrition may occur in which attrition is associated with the treatment versus control group status of the participant. This form of attrition may undermine the success of the randomization procedure (link B in Figure 6.1), potentially jeopardizing the researcher's ability to infer the treatment had a causal effect, even for the restricted population that would complete the full experimental protocol including posttest measurement. Below, we consider methods of detecting participant characteristics associated with each of these forms of attrition and methods for correcting estimates of treatment effects for these effects of attrition.

Identifying Sources of General and Differential Attrition

Cook and Campbell (1979) strongly emphasized the importance of pretest measures in attempting to understand and model the effects of attrition in randomized experiments. They advocate collecting pretest measures from all participants on important variables that are likely to be related to the outcome measure of interest, most notably the pretest on the same measure. Given such data, they recommend a two-step strategy originally proposed by Jurs and Glass (1971) to detect covariates associated with general and differential attrition. Here, we recommend a modification of the second step that uses an improved statistical estimation procedure. Both steps must be passed before tentatively concluding that attrition does not appear to have an appreciable effect on the results.

The first step is to compare the proportion of participants subject to attrition in the treatment and control groups. If these proportions differ, then differential attrition may be interpreted as a treatment effect. The interpretation of all other measured outcome variables becomes potentially problematic (Sackett & Gent, 1979).

The second step is to conduct a series of logistic regression analyses⁸ to predict attrition status (completer vs. "attriter"). Separately for each measured pretest variable, attrition status is predicted by treatment assignment (treatment vs. control), the pretest variable, and the interaction of the pretest variable and treatment assignment. Following Aiken and West (1991) and West, Aiken, and Krull (1996), the pretest variable should be centered (i.e., in deviation score form) and the treatment assignment should be effect coded. In these analyses, the first-order effect of the pretest variable represents general attrition (threat to external validity), and the interaction of treatment condition and pretest variable represents differential attrition (threat to internal validity). The goal of these analyses is to identify all possible participant characteristics (assessed prior to treatment) that may be associated with attrition in the population. These characteristics are then used in statistical procedures that attempt to adjust the estimates of treatment effects for the identified predictors of attrition.

Two major problems arise in the use of either the original Jurs and Glass (1971) procedure or the modification we proposed above. First, attrition may be related to what Rosenbaum (1995) has termed *hidden variables*, participant characteristics that are *not* assessed at pretest. Conceptually, the potential effects of hidden variables exactly parallel those of measured covariates. If the hidden variable is equally related to attrition status in the treatment and control groups, generalization of the causal effect to the population of interest may be limited. If the hidden variables have differential relationships to attrition status in the treatment and control groups, then the estimate of the causal effect may be biased. The magnitude of the bias in each case also depends on the magnitude of the relationship between the hidden variables and the outcome variable. The central problem is that hidden variables are unmeasured—the information necessary for a proper adjustment of the treatment effect is not available.

Second, our logistic regression modification of the Jurs and Glass procedure uses null hypothesis significance tests to identify pretest characteristics that potentially relate to attrition. The success of this procedure assumes that the specification of the logistic regression equation as a linear model is correct⁹ and that there is sufficient statistical power to detect any attrition-related effects of appreciable magnitude. The issue of sufficient statistical power has, in particular, received considerable discussion in the literature, given that many randomized trials do not have large sample sizes. For example, Hansen, Collins, Malotte, Johnson, and Fielding (1985) have called for using less conservative

levels of Type I error (e.g., $\alpha = .25$) in testing attrition-related hypotheses; Tebes, Snow, and Arthur (1992) have proposed setting the Type I error rate so that the ratio of the Type II to Type I errors will approximate 4:1 (viz. $\beta = .20$ and $\alpha = .05$). This literature serves as a strong reminder that the central goal of the Jurs and Glass procedure is to identify *all* possible sources of general and differential attrition that should be retained for further investigation. Even if a large number of tests is performed, the significance level associated with each effect should *not* be corrected for the experimentwise error rate (alpha inflation). Such correction is contradictory to the basic goal of the procedure.

Statistical Adjustments for Attrition

Continuous Outcome Variables

Traditionally, statistical and methodological techniques have focused on providing an appropriate adjustment for differential attrition (see Foster & Bickman, 1996, for a review of techniques). One simple statistical approach is to include all pretest variables identified using the second step of the Jurs and Glass procedure as covariates in an analysis of covariance, now focusing on effect of the treatment on the *outcome variable* of interest. This procedure removes the bias in the estimate of the treatment effect due to the measured covariates. If good estimates of the reliabilities of the pretest variables are available, further improvements may be made in the estimate of the treatment effect using procedures that correct for unreliability in the covariates (Huitema, 1980; West, Biesanz, & Pitts, in press). More complex statistical weighting procedures that use *propensity scores* (i.e., the estimated probability that the participant will respond on the outcome variable based on available information from the participant's pretest responses and treatment condition; see Rubin and Thomas [1996]) have also provided highly accurate estimates of treatment effects in some evaluations (McGuigan et al., 1997).¹⁰

Alternatively, researchers may take approaches that attempt to augment the data available for the outcome variable. One strategy that can be used in some large studies is to try to understand the reasons for attrition (Graham et al., 1994), identifying classes of attriters that theoretically might be expected to have different outcomes from those observed for the completers. For each class of attriter, the researcher would take a random subsample and make heroic (and expensive; see Graham & Donaldson, 1993) efforts to locate and measure as close to 100% of these individuals on the outcome variable as possible. To illustrate, in our example of the substance abuse prevention program, students may not be measured on the outcome variable because (a) they were absent from

school, (b) they have transferred to another school, (c) they have dropped out of school, or (d) they refuse to participate. A random subsample of attriters from each of these classes could be selected. Special efforts could then be made to locate them and provide appropriate incentives so that each of the attriters in these four subsamples could be remeasured. The primary advantages of this strategy are that it focuses the limited resources of the study and theoretically can provide an unbiased estimate of the treatment effect if 100% remeasurement is achieved. In the absence of full remeasurement, estimates of the treatment effect make the assumption that an achieved sample represents a random sample of the class of attriters.¹¹

A second strategy is to identify another data source that could provide information on the outcome variable. For example, peer reports and parent reports of the child's substance use might be collected for the attriters and for a random sample of the completers on which the participant's own report was available. Assuming that data are missing at random, this information would then provide a strong basis for imputing the attriters' level on the outcome variable of interest.

Modern missing data theory (Little & Rubin, 1987; Schafer, 1997) provides several useful approaches for imputing data when data are missing at random. These methods simultaneously address both general attrition and differential attrition. One approach is the EM algorithm (Dempster, Laird, & Rubin, 1977; Little & Rubin, 1987), which produces a very good estimate of the covariance matrix between the measured covariates, treatment group assignment, and the outcome, from which the treatment effect may be estimated. A second approach is multiple imputation (Rubin, 1987, 1996; Schafer, 1997), in which a regression equation is used to predict the specific value for the outcome variable for participants with missing data. As part of the multiple imputation process, random error is added to these predicted values so that the variance of variables having missing data will not be affected by the procedure. Because the random error will vary with each imputation, several data sets are produced. Each data set is analyzed separately, and the mean of the estimates of the treatment effect across these data sets is taken as the best value. Although both the EM algorithm and the multiple imputation procedures provide appropriate corrections for general attrition and differential attrition when data are missing at random, both typically assume that the treatment effect is constant (Figure 6.2A). If specific Treatment \times Covariate interactions are expected, as might be the case in Figure 6.2B, alternative imputation procedures that build these interactions into the model being tested may need to be undertaken (Arbuckle, 1996). These model-based procedures make the strong assumption that the model being estimated is correct and that the same model can be used to characterize both participants with complete data and participants with incomplete data. Finally, sensitivity analyses may be conducted in which the effect of assuming a range of

covariate-outcome and treatment-outcome relationships for participants with incomplete data may be explored (Little & Rubin, 1987; Rubin, 1977).

None of the approaches presented above has addressed the issue that attrition may be associated with a hidden variable that was not assessed at pretest. Drawing on ideas proposed by Rosenbaum (1995) in another context, we propose a form of sensitivity analysis that can be used to suggest lower and upper bounds for the estimates of the magnitude of the treatment effect. The researcher would begin by identifying the largest standardized difference observed on *any* of the covariates at pretest between participants assigned to the treatment and control groups, $d = (\bar{X}_t - \bar{X}_c)/s$. Given that a sizable number of the most important covariates have been measured at pretest, d can be taken as a reasonable estimate of the maximum plausible standardized difference between the pretest means in the treatment and control groups for the unmeasured variable. The researcher would then calculate the pretest-posttest correlation r for the outcome variable of interest, measured within the control group. This value provides an estimate of the maximum correlation between the unmeasured pretest variable and the outcome variable. The product $d \times r$ then provides a reasonable upper-bound estimate of the maximum difference in the standardized effect on the outcome variable that could be expected between attriters and completers as the result of a hidden variable. The product $d \times r$ would then be multiplied by the ratio of the proportion of attriters to the proportion of completers in the sample to arrive at a value of an adjustment factor for the estimate of the standardized effect size. This adjustment factor could be added to and subtracted from the best estimate of the treatment effect to provide reasonable upper and lower bounds for the effect. This procedure assumes that the treatment effect is constant and that missingness is conditioned on the single hidden variable. Nonetheless, the values chosen for d and r typically can be considered to be maxima, so that this procedure can be expected to provide plausible upper and lower bounds for treatment effects in most real-world contexts.

Dichotomous Outcome Variable

Paralleling the approach taken by Robins (1989) to the issue of treatment noncompliance, Shadish, Hu, Glaser, Kownacki, and Wong (1998) have developed an alternative approach that attempts to bracket the estimate of the treatment effect. This approach is applicable in research contexts with two treatment conditions and dichotomous outcomes. Instead of attempting to generate a unique treatment effect estimate, Shadish et al. enumerate the distribution of potential treatment effects under all possible combinations of outcomes for treatment and control group attriters. Then, if the user can specify information about the distribution of outcomes for the attriters, a probability can be calcu-

lated representing the odds of obtaining a treatment effect below a specified threshold. To illustrate, consider a smoking cessation experiment in which participants are randomly assigned to treatment and control groups. Assume that a complete-case analysis (listwise deletion) of participants whose smoking status can be observed at posttest reveals a significant effect of treatment, but that the smoking status of a substantial proportion of participants cannot be reassessed. If the researcher makes the assumption that .80 of the dropouts will resume smoking, then the Shadish et al. approach can be used to determine the probability of obtaining a nonsignificant effect. In some research contexts, past literature can provide a basis for assumptions about the plausible bounds on the unobserved outcomes of the attriters. In the absence of information that permits a plausible assumption, an attrition analysis plot may be generated that graphically depicts the distribution of probabilities under all possible combinations of assumptions regarding treatment and control group attrition.

Summary

Different statistical approaches may be taken to attrition when the outcome variable is continuous versus dichotomous. For continuous outcome variables, the methods identify measured covariates that are associated with attrition and attempt to provide an appropriate adjustment of the point estimate of the treatment effect. These methods assume that there are no hidden variables (data are MAR), the covariates have been measured without error, and the relationship between the covariates and the outcome variable has been specified correctly. They also assume that participants subject to attrition received the full active or control treatment to which they were assigned.¹² In contrast, the Shadish et al. approach develops bounds for the treatment effect that reflect the influence of attrition. Reflecting the simplification possible with dichotomous outcomes, the Shadish et al. procedure is not dependent on the set of assumptions noted above; rather, this procedure simply generates either (a) the distribution of all mathematically possible outcomes that are consistent with the observed data or (b) a more constrained set of plausible outcomes that are consistent with both the data and available empirical knowledge.

THE CAUSAL EFFECT IN THE ACHIEVED SAMPLE VERSUS THE POPULATION

As we observed in the introduction, the selection of a random sample from a defined population is almost never achieved in the context of randomized trials. Many trials do not use a formal sampling procedure to select participants from their population of interest, violating Stage A of the formal two-stage statistical

model. Those trials that do employ some form of random sampling procedure typically are unable to enroll a majority of the sampled participants into the trial (e.g., Vinokur et al., 1991; Wolchik, Sandler, West, & Anderson, 1997). Thus, even if the problems of treatment noncompliance and attrition were not present, there is still no guarantee that the average effect size observed in the trial will necessarily characterize the full population. To the extent that the actual achieved sample is not representative, there is a potential for bias in the estimate of the average treatment effect in the population. The extent of the actual bias in the treatment effect estimate will depend on the *multiplicative product* of two features of the data.

1. *Nonrepresentative sample.* The greater the extent to which the distribution of participant characteristics in the sample does not reflect the distribution in the population, the greater the risk of bias.
2. *Nonconstant treatment effect.* The greater the extent to which the treatment effect is nonconstant (i.e., treatment conditions interact with participant characteristics such as gender or initial health status), the greater the risk of bias.

When a sample is nonrepresentative with respect to those specific participant characteristics that are associated with the nonconstant treatment effect, the estimate of the average causal effect of the treatment in the population may be substantially higher or lower than the true value.

To provide an estimate of the potential bias, the population and the achieved sample need to be compared. Many populations of interest to researchers in a community (e.g., schools, courts, unemployment offices, hospitals) maintain extensive records on individuals.¹³ Alternatively, a community survey or census that includes information about the population of interest may be available or may be undertaken. In such cases, the demographic and other relevant information available on the full population of interest may be compared with the same information collected from the achieved sample.¹⁴ Paralleling the Jurs and Glass (1971) procedure for attrition, the purpose of these comparisons is to identify any characteristics of participants in the sample that may not be representative of the values of these same characteristics in the population. When discrepancies are found, weighted statistical procedures (e.g., weighted ordinary least squares regression; Winship & Radbill, 1994) may be used to adjust the estimates of the treatment effect so that they more closely reflect the actual distribution of cases in the population. Such procedures promise estimates that may more closely reflect the true treatment effect in the population. These procedures make the strong assumption that the data are missing at random: Following adjustment for all detected sources of nonrepresentativeness, any other undetected sources of nonrepresentativeness are expected to have ignorable effects on the estimate of the treatment effect.

There is a second, more focused basis on which treatment effects may be adjusted. If the researchers carefully record the process of recruiting participants into the experiment, a careful study can identify those steps in the recruitment process at which selection seems to occur. For example, consider a randomized trial of a parenting program for recently divorced mothers with young children (Wolchik et al., 1997). The trial began with a random sample of mothers with at least one child 8 to 15 years of age, selected from county divorce records. These participants were then carefully tracked through seven separate steps of the recruitment process. Only two of the seven steps appeared to be associated with substantial selection bias. First, families who were more difficult to find during the initial location process (i.e., *hard to locates*: they had no listed address or telephone) had children with worse scores on an initial measure of mental health problems. Second, families who declined participation in the trial at the final recruitment step (*late decliners* who dropped out prior to randomization following a 15-minute in-home description of the protocol of the trial) had children with better scores on an initial measure of mental health problems. Preliminary analyses showed an interaction between initial (baseline) level of mental health problems and treatment condition (program versus control) on the primary outcome variable—children's level of mental health problems at posttest. Given the potential for bias in the estimation of the treatment effects, weighted regression analyses were conducted that adjusted for the underrepresentation of (a) the hard-to-locate and (b) late-decliner subpopulations in the achieved sample. The results of these analyses suggested that the initial unadjusted estimate of the average treatment effect may have been approximately 5% too low. This adjusted value provides an estimate of the treatment effect that would have been obtained in the full population of divorced families (i.e., if all divorced families could have been enrolled in the trial). This adjusted estimate depends on two important assumptions: (a) The achieved sample of hard-to-locate families is representative of the subpopulation of hard-to-locate families¹⁵ and (b) the same treatment effect characterizes participants and nonparticipants in the randomized trial after conditioning on participant characteristics associated with selection.

A third design-based method for studying selection into randomized trials has been proposed by Braver and Smith (1996). Braver and Smith note that, following Cook and Campbell (1979), randomization frequently is offered only to participants who would accept *any* of the treatment conditions (i.e., they are indifferent to the treatment conditions). If the available treatment conditions differ greatly in their attractiveness to participants, the resulting sample will likely be very unrepresentative of the full population of interest. Braver and Smith suggest that a more elaborate research design, which they term the *combined modified design*, be used in such contexts. In this design, the sample is randomly assigned to one of two sub-experiments. The first sub-experiment

follows the traditional procedure: Participants are randomly assigned to treatment and control conditions after they have agreed to participate in any of the treatment conditions. This sub-experiment maximizes internal validity, but at a potential cost in generalization to the full population. In the second sub-experiment, participants are randomly assigned to conditions in a randomized invitation (encouragement) design. The participants would hear the description of only one of the programs and would be encouraged to participate. Statistical techniques described in our earlier section on treatment noncompliance could then be used to provide estimates of the causal effect. The second sub-experiment allows clear generalization of the estimate of the treatment effect either to the population of compliers (Angrist et al., 1996) or the full population (Robins, 1989), depending on the statistical procedures that are used. The cost of this generalization is that the estimates of the causal effect are dependent on stringent assumptions; however, by combining information from the two sub-experiments, the researcher has the potential to maximize simultaneously both the internal and external validity of the results, assuming that the attractiveness of the treatment condition to participants is the major limitation on generalization. If other aspects of the randomized trial, such as the burden of completing the measurements or concern about exposure to unproven treatments, are important sources of selection into the randomized trial, generalization of the results to the population may still be limited.

CONCLUSION

In this chapter, we have presented a number of modern methods that help answer questions about the effects of a treatment on a defined population of participants. We first considered participant noncompliance, examining both classic and modern statistical approaches that under certain circumstances allow unbiased estimates of treatment effects. We then considered approaches that help detect sources of bias that may result from participant attrition and methods for adjusting treatment effects that help remove these sources of bias. Finally, we considered methods of identifying and correcting for sources of bias associated with participant selection into the randomized trial. Each of these techniques is based on important assumptions. If these assumptions can be met—and they are often stringent—then unbiased quantitative estimates of the causal effect in the population can be calculated. Such quantitative estimates potentially allow researchers and policymakers to make judgments of whether the treatment or the program would have meaningful effects if implemented.

At the same time, the present approaches are not a panacea. Campbell and his associates (Campbell, 1957; Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish et al., in press) as well as Cronbach and his associates (Cronbach, 1982; Cronbach et al., 1980) have repeatedly reminded researchers of the diffi-

culty of generalizing the results of a single randomized trial, even one for which an unbiased estimate of the causal effect is available. The statistical methods presented in this chapter are based on Rubin's causal model in which both the treatments and the outcome measure are assumed to be *precisely* defined (see Holland, 1986; Rubin, 1986). Given such precise definition, Rubin's causal model permits the causal effect inferred from the results of the randomized trial to be precisely generalized to the defined participant population of interest. Yet, as noted by Campbell, Cronbach, and their associates (see also Cook, 1993; Reichardt, Chapter 5, this volume), the observed level of the outcome variable is dependent on the (a) units (participants), (b) treatment, (c) observations (measures), (d) setting, and (e) time in which the data are collected. Unless all these conditions are comparable in the randomized trial and in the actual implementation of the treatment, there is no guarantee that the estimate of the causal effect will generalize precisely to the population of interest.

Of importance, Shadish et al. (in press) note that although random selection of participants (units) can sometimes be approximated, random selection of treatments, observations (measures), or settings from defined populations of potential treatments, observations, and settings is virtually never undertaken. If researchers and policymakers wish to generalize the results of a randomized trial to a different population, to a modified treatment, to a new outcome measure, and/or to a new setting, they cannot rely on the formal statistical model based on random selection. In practice, treatments evolve and change when they are delivered by practitioners (Mayer & Davidson, in press; Sechrest, West, Phillips, Redner, & Yeaton, 1979). Settings also can change, as when a new medical treatment is now delivered in the setting of an HMO rather than a prestigious university research hospital. Such changes have a real potential for modifying the magnitude of causal effects, yet Rubin's causal model contains no formal mechanism for addressing these issues of generalization. These issues underlie the potential difficulty in generalizing the results of efficacy trials in which the treatment is tested under optimal experimental conditions to effectiveness trials in which the treatment is tested under the more realistic conditions in which the treatment typically is delivered in practice.

Cook (1993) and Shadish et al. (in press) have recently begun to articulate practical principles for thinking about these issues of generalization. Their principles represent a systematization of some of the insights of Donald Campbell and Lee Cronbach, their coworkers, and students about external validity and construct validity. In brief, these principles may be summarized as follows (see Cook, 1993, for a full presentation).

1. *Proximal similarity.* Use specific units, treatments, observations, and settings in the randomized trial that correspond as closely as possible to units, treatments, observations, and settings of interest in the population.

2. *Heterogeneous irrelevancies.* Use multiple instances of units, treatments, observations, and settings in the randomized experiment that are heterogeneous with respect to aspects of these dimensions that are theoretically expected to be irrelevant to the treatment-outcome relationship in the population.
3. *Discriminant validity.* Use a treatment that produces changes in the intended causal agent (conceptual independent variable) and not other causal agents. Use an outcome measure that assesses the intended conceptual dependent variable, but not other conceptual variables.
4. *Causal explanation.* To the extent that the researchers can support a causal explanation of the findings and rule out competing explanations, the likelihood of generalization to the population is increased.
5. *Empirical interpolation and extrapolation.* To the extent that generalization involves interpolation within (rather than extrapolation beyond) the range of values of units, treatments, observations, and settings that have been studied, the likelihood of generalization is increased.

Shadish et al. (in press) have outlined how these principles can be used as a basis for generalizing the results of a single randomized trial as well as the results of a meta-analysis of an entire research literature to a population of units, treatments, observations, and settings of interest. These principles are very useful in making statements about the likely presence or absence of a treatment effect when the treatment is actually implemented in the population of interest. Note, however, that these principles do not presently include any formal mechanism for estimating the *magnitude* of the treatment effect that would be likely to occur when the treatment is implemented in the population of interest. The continued development of these principles of causal generalization, combined with the improved estimates of effect size available through the use of the statistical procedures described in this chapter, offers great promise of helping researchers develop an understanding of when the causal effects found in randomized trials can and cannot be meaningfully generalized to the units, treatments, observations, and settings of interest.

NOTES

1. All estimates and standard errors are unbiased. Even when data are missing completely at random, however, there is a loss of statistical power relative to complete data (see Graham, Hofer, & Piccinin, 1994).

2. One common variant of the TR approach in psychological research is that only those participants who are assigned to a treatment condition who actually receive that treatment are included in the analysis; all other cases are discarded from the analysis. In the example in Figure 6.3A, participants 1 and 2 would be thrown out, leaving the comparison of the mean of the responses of participants 3-6 in the treatment group ($\bar{X}_1 = 5.75$)

with the mean of responses of participants 7-12 in the no treatment comparison group ($\bar{X}_c = 4.0$). As can be seen, this estimate of the treatment effect (1.75) is too high. Procedures that discard participants suffer from similar problems to the analysis by treatment received discussed in this section.

3. Given the potential for bias, researchers may wonder whether their sample of participants includes defiers. Do these troublesome people actually exist, or are they postulated only to ensure that the set of categories of compliance is exhaustive? Masling (1966; see also Silverman, 1965) speculated that some participants in psychological experiments might resist experimental instructions in order to avoid giving the impression that the experimenter could control their behavior. Christensen (1977) showed in a pair of experiments that participants who had taken part in a previous highly manipulative experiment resisted manipulative attempts in a future experiment. A cautious interpretation of these findings suggests that defiers might possibly exist in experiments with highly coercive treatment *and* control conditions. Thus, we conclude that defiers are unlikely to exist in nearly all randomized field experiments. Of course, in some cases a researcher concerned about defiers can ensure monotonicity by making treatment unavailable to controls.

4. Two issues are particularly important in evaluating the quality of the compliance measure. First, standard measures assume that each compliance opportunity is equivalent. Missing particular sessions in a standard multisession treatment program or missing particular doses in a pharmaceutical regimen may have implications for the strength of the treatment that the participant receives. Second, Meier (1991) notes some of the many ways in which self-presentational issues may distort measures of compliance. These include inaccurate self-reports and changes in compliance behavior in anticipation of scheduled measurement of participant knowledge, behavior, or physiological state (e.g., blood or urine tests).

5. Holland's ALICE model requires a stronger form of the exclusion restriction assumption (see Angrist et al., 1996, for a discussion of the strong versus weak forms of the exclusion restriction).

6. Dawes (1979) provides evidence that linear models can often provide reasonable approximations to other monotonic functions. Holland's assumption of a linear relationship between the measure of compliance and the outcome, however, can be expected to break down when there is a wide range of values on the compliance measure. To illustrate, linearity implies that a particularly dedicated student would benefit as much from her 73rd hour of studying as from her 2nd.

7. This assumption implies that all participants will respond identically to the same treatment dosage, regardless of how they would have responded if given no treatment. Models in which this assumption is relaxed can be estimated; however, the resulting estimate of the magnitude of the treatment effect is associated with considerable uncertainty.

8. Jurs and Glass (1971) originally recommended a two-way analysis of variance with treatment condition and attrition status as the factors and the pretest measure as the outcome. The logistic regression analysis provides improved estimates as the proportion of participants who drop out becomes increasingly discrepant from .50. Logistic regression can also be used to provide propensity scores for use in weighting procedures to correct for nonresponse (McGuigan et al., 1997). The issues in specifying functional form in

the logistic regression remind us that randomization leads to the expectation that treatment assignment will be independent of all participant background characteristics—a much stronger condition than a lack of linear relationship.

9. Nonlinear relationships and interactions potentially can be detected through graphical examination of the relationship between the pretest variable and the outcome variable (the log odds of attrition) within each group separately.

10. Heckman's (1979) sample selection model has sometimes been recommended to adjust for problems of attrition; however, when its assumptions are not met, this approach has been shown to produce severely biased estimates of treatment effects in both simulation studies and actual evaluations (e.g., McGuigan et al., 1997; Stolzenberg & Relles, 1990).

11. Large standard errors may also result if resources permit relocation and remeasurement of only a small number of attriters in each class.

12. The mechanisms producing treatment noncompliance and attrition are likely to be different. Consequently, different adjustment strategies are likely to be needed for the two problems.

13. In many research contexts (e.g., children who have an alcoholic parent), the population of interest is "hidden" and cannot be enumerated. Different recruitment procedures may produce samples with strikingly different participant characteristics (Chassin, Barrera, Bech, & Kossak-Fuller, 1992). Bryan (1998) discusses issues in sampling from such hidden populations.

14. Unfortunately, in many research areas such information traditionally has not been sought out on the population, nor has useful descriptive information been collected on participants in the sample (Moskowitz, 1993).

15. Wolchik et al. (1997) conducted a substudy on a random sample of the hard-to-locate families in the population (e.g., no listed phone or address). Intensive location procedures (e.g., contacting relatives and former neighbors) succeeded in locating more than 90% of this subsample. No differences were detected between the hard-to-locate families found through normal versus intensive location procedures.

REFERENCES

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Albert, J. M., & Demets, D. L. (1994). On a model-based approach to estimating efficacy in clinical trials. *Statistics in Medicine*, *13*, 2323-2335.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444-455.
- 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. 243-277). Mahwah, NJ: Erlbaum.
- Balke, A., & Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In R. Lopez de Mantaras & D. Poole (Eds.), *Proceedings of*

- the Tenth Conference on Uncertainty in Artificial Intelligence* (pp. 46-54). San Francisco: Morgan Kauffman.
- Balke, A., & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92, 1171-1176.
- Biglan, A., Hood, D., Borzovsky, P., Ochs, L., Ary, D., & Black, C. (1991). Subject attrition in prevention research. In C. G. Luekfeld & W. Bukowski (Eds.), *Drug abuse prevention intervention research: Methodological issues* (NIDA Research Monograph #107) (pp. 213-234). Rockville, MD: National Institute on Drug Abuse.
- Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, 8, 225-246.
- Boruch, R. F. (1997). *Randomized experiments for planning and evaluation*. Thousand Oaks, CA: Sage.
- Braver, S. L., & Smith, M. C. (1996). Maximizing both external and internal validity in longitudinal true experiments with voluntary treatments: The "combined modified" design. *Evaluation and Program Planning*, 19, 287-300.
- Brewer, M. B. (1976). Randomized invitations: One solution to the problem of voluntary treatment selection in program evaluation research. *Social Science Research*, 5, 315-323.
- Bryan, A. D. (1998). *Strategies for reaching hidden populations: A review of methods*. Unpublished manuscript, Department of Psychology, University of Colorado, Boulder.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297-312.
- 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* (New Directions for Program Evaluation, no. 31, pp. 67-77). San Francisco: Jossey-Bass.
- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- 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.
- Christensen, L. (1977). The negative subject: Myth, reality, or a prior experimental experience effect? *Journal of Personality and Social Psychology*, 35, 392-400.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. *New Directions for Program Evaluation*, 37, 39-81.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

- Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements*. San Francisco: Jossey-Bass.
- Dawes, R. M. (1979). The robust beauty of improper linear models. *American Psychologist*, *34*, 571-582.
- Dempster, A. P., Laird, N., & Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *B39*, 1-38.
- Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, *20*, 115-147.
- Efron, B., & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, *86*, 9-17.
- Foster, E. M., & Bickman, L. (1996). An evaluator's guide to detecting attrition problems. *Evaluation Review*, *20*, 695-723.
- Gochman, D. S. (Ed.). (1997). *Handbook of health behavior research* (Vol. 2). New York: Plenum.
- Goetghebeur, E., & Molenberghs, G. (1996). Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association*, *91*, 928-934.
- Goetghebeur, E., & Shapiro, S. H. (1996). Analysing non-compliance in clinical trials: Ethical imperative or mission impossible? *Statistics in Medicine*, *15*, 2813-2826.
- Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and follow-up data. *Journal of Applied Psychology*, *78*, 119-128.
- Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 325-366). Washington, DC: American Psychological Association.
- Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. In L. M. Collins & L. A. Seitz (Eds.), *Advances in data analysis for prevention intervention research* (NIH Publication No. 94-3599) (pp. 13-63). Rockville, MD: National Institute on Drug Abuse.
- 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.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153-161.
- 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.
- Huitema, B. E. (1980). *The analysis of covariance and alternatives*. New York: Wiley.