10

Practical Problems 2: Treatment Implementation and Attrition

Im·ple·ment: (ĭm´plə-mənt) v. *tr.* [Middle English *supplementary payment,* from Old French *emplement, act of filling,* from Late Latin *implementum,* from Latin *implere.*] 1. To put into practical effect; carry out: implement the new procedures. 2. To supply with implements. **imple·men·tation** n. **imple·menter** n.

At·tri·tion (ə-trĭsh´ən): n. [Middle English *attricioun, regret, breaking,* from Old French *attrition, abrasion,* from Late Latin *attritio,* attrition- *act of rubbing against,* from Latin *attritus,* past participle of *atterere, to rub against:* ad-, *against.*] 1. A rubbing away or wearing down by friction. 2. A gradual diminution in number or strength because of constant stress. 3. A gradual, natural reduction in membership or personnel, as through retirement, resignation, or death.

IN AN ideal world, those who are assigned to an intervention would receive it properly and would be measured on all outcomes, and those assigned to the comparison condition would also receive the proper treatment (or receive no treatment, according to the design) and be measured. But field experiments are rarely ideal. The intervention may not be implemented properly, fully, or even at all. Sometimes, those assigned to the intervention group refuse the treatment either before or after the experiment starts; and sometimes, the controls insist on receiving the intervention. Other people drop out of the study entirely so that no outcomes are ever observed on them, whether or not they received the intervention. All these problems fall under two headings, treatment implementation and attrition. These two problems threaten the very reason for doing an experiment:

to get a good estimate of a treatment effect. In this chapter, we describe both design and statistical approaches to both these problems.

PROBLEMS RELATED TO TREATMENT IMPLEMENTATION

In a randomized experiment that studied a program to increase medication adherence for newly diagnosed patients with high blood pressure, each patient in the treatment group was given a form on which to record his or her medication intake as an aid to improve adherence (Saunders, Irwig, Gear, & Ramushu, 1991). After 6 months, only 28% could even produce the form, and only 40% knew how to fill it out correctly. Another study found that only 30–40% of patients took their medications within 2 hours of the time of day they were supposed to be taken, even with weekly reminders (Rudd, Ahmed, Zachary, Barton, & Bonduelle, 1990). In the Minneapolis Spouse Abuse Experiment (Berk et al., 1988), 18% of participants received a treatment other than the one to which they were assigned. Bloom (1990) found that about 3% of the unemployed workers who were assigned to the no-treatment condition crossed over into one of the treatment conditions. In a psychotherapy outcome study, the same therapists administered both treatments, and one therapist used extensive techniques from the behavioral condition when he was doing therapy with clients in the eclectic condition (Kazdin, 1992).

These examples point to three related problems—failure to get the full intervention, crossing over to get a different treatment, and treatment diffusion—that are common in field experimentation. Indeed, few experiments achieve full implementation of the intervention, in which every participant in each condition is offered the intervention, fully receives the intervention, and fully complies with the intervention to which they were assigned and no other intervention. Knowledge of treatment implementation helps characterize the intervention and its context, facilitates the detection of problems in the intervention, and helps distinguish between the effects of assignment and the effects of the intervention. This section discusses practical methods for assessing and inducing implementation and then presents statistical analyses that account for levels of implementation.

Inducing and Measuring Implementation

Experiments benefit from making sure treatment is implemented as intended (induction), and from having very specific information about the extent to which the intervention is actually delivered and then received and implemented by the recipient. Such information helps to ensure that the intervention was actually manipulated, to detect and remedy problems with the intervention before they progress too

far, to describe the nature of the intervention, to explore not just whether the intervention works but how it works, and to examine covariation between intervention and outcome. All these issues bear directly on the cause-effect relationship that is the central justification for doing a randomized experiment in the first place.

The Components of Implementation

Treatment implementation is not just one thing but rather is a multifaceted process that includes **treatment delivery, treatment receipt,** and **treatment adherence** (Lichstein, Riedel, & Grieve, 1994). To use a simple medical example, *delivery* refers to whether a physician writes a prescription for a patient, *receipt* refers to whether the patient gets the prescription and has it filled, and *adherence* refers to whether the patient takes the prescribed medication and does so according to the instructions. For some cases, these three processes blend together. For example, delivery, receipt, and adherence often blend together completely for a surgical intervention in which the delivery and receipt of, say, an incision are two sides of the same action. In other cases, these three processes could be broken down further, especially in the kind of complex social interventions that occur in, say, whole school reform, job training, or programs such as Head Start.

Adherence is what most researchers probably mean when they talk about treatment implementation. But adherence clearly depends on delivery and receipt in most of the interventions studied in social experiments. Consequently, the researcher should consider doing things to increase the chances that delivery, receipt, and adherence will occur and to measure how much delivery, receipt, and adherence actually happened in the experiment. Here are some examples.

Inducing and Measuring Treatment Delivery. The odds that an intervention will be delivered can be increased by using treatment manuals, by training service providers, by giving verbal reminders to those providers to include all treatment procedures, by providing on-the-spot instructions to them during treatment, or by administering treatment by videotape or audiotape. In addition, treatments may be delivered with less integrity when they are more complex, burdensome, of long duration, inconvenient, or expensive or when they require the recipient to alter his or her lifestyle; so reducing all these problems when possible can increase service delivery. Delivery can be measured in staff meetings by supervisors of service providers or by reviewing or even formally scoring tapes of the sessions at which service is provided. Assessing differential delivery (e.g., that the treatment *excludes* key components of a comparison treatment) is important when many treatments are being compared to each other. For example, the NIMH Treatment of Depression Collaborative Research Program used ratings of psychotherapy sessions to show that therapists in three different therapy conditions performed those behaviors appropriate to the condition to which they were assigned more often than they performed the behaviors appropriate to the other conditions (Hill, O'Grady, & Elkin, 1992).

Inducing and Measuring Treatment Receipt. Ways to improve treatment receipt include giving written handouts to recipients that summarize key treatment points, using established communication strategies such as repetition of the message or making the treatment deliverer appear expert or attractive, questioning the recipient about key treatment features to induce cognitive processing of treatment, and having recipients keep logs of their treatment-related activities. In many cases failure of treatment receipt is due to failure of communication between the deliverer and the recipient if the provider is a poor communicator or if the recipient is poorly motivated or inattentive. Receipt can be measured using manipulation checks, written tests of change in recipients' knowledge related to treatment, talk-aloud assessments of the recipients' experience during treatment, monitoring physiological changes that the treatment should induce when it is received, or asking the recipients if they are confident applying treatment skills.

Inducing and Measuring Treatment Adherence. Adherence is reduced when treatment participants lack the time to carry out the treatment, forget to do it, are unsure of the correct treatment procedures, are disappointed by initial results from trying treatment, lack access to an appropriate setting in which to carry out the treatment, or simply lose motivation to change (Lichstein, 1988). Each of these suggests strategies for improving adherence; for example, by assigning the recipient written homework to do and return, by using family members to encourage adherence, by making available physical aids such as tape recordings or motivational cards to guide practice of the skills taught in the intervention, or by giving reinforcements such as raffle tickets for demonstrated adherence. Adherence can be measured by interviewing the recipient and other informants about the recipient's treatment-relevant activities outside of treatment or by using biological assays when adherence to the treatment would result in a detectable biological change, such as the presence of a medication in the blood or the absence of nicotine residuals in saliva for a stop-smoking treatment. Adherence has been particularly heavily studied in medicine; entire monographs (Cramer & Spilker, 1991; Haynes, Taylor, & Sackett, 1979; Okene, 1990; Sackett & Haynes, 1976; Schmidt & Leppik, 1988; Shumaker & Rejeski, 2000) are devoted to the topic, and even a journal (the *Journal of Compliance in Health Care*) dedicated to adherence existed for 4 years (1986–1989).

Overlooked Targets for Implementation Assessments

Three kinds of implementation assessments are often overlooked in experiments. First, researchers often forget to assess the *extra-study* treatments that participants are getting while they are in an experiment. In AIDS research, for example, patients who are not assigned to the treatment they prefer often seek their preferred treatment elsewhere, all the while remaining in the treatment arm to which they were assigned (Marshall, 1989). Some AIDS patients are enrolled in more than one experiment simultaneously (Ellenberg et al., 1992). Similarly, in studies of the

effects of social programs to improve pregnancy outcome, participants often sought and used other programs to improve the health of their babies (Shadish & Reis, 1984).

Second, researchers often forget to assess the treatments received by those who are assigned to a no-treatment control condition, for it is naive to think that they experience nothing between a pretest and posttest. The clearest examples occur in medical trials in which some patients must be taken off placebo medication and treated actively for ethical reasons. Often, however, those assigned to control groups actively seek treatment elsewhere. For example, Margraf et al. (1991) questioned patients and physicians in a randomized, double-blind experiment that compared two medications for treatment of panic disorder with a placebo and found that the great majority of both patients and physicians could guess accurately whether they received active medication or the placebo (see also Greenberg, Bornstein, Greenberg, & Fisher, 1992). Given the extreme distress that panic disorder patients often experience, some patients assigned the placebo probably sought and received treatment elsewhere.

Third, researchers often forget to assess the unplanned things that service providers do in treatment. For example, those who administer treatment may depart from the protocol by adding new components based on their experience of what works, making a treatment more powerful (Scott & Sechrest, 1989; Sechrest, West, Phillips, Redner, & Yeaton, 1979; Yeaton & Sechrest, 1981). By definition, planning to measure such unplanned deviations is difficult. So the experimenter may need some capacity for discovery. Qualitative methods such as participant observation or open-ended interviews with providers and recipients can provide that capacity.

Assessing Program Models

To this point, we have been speaking of implementation issues as if they are limited to the intended treatment. We have done so in order to emphasize the role that these matters play in improving the construct validity of the treatment and in allowing the use of the statistical analyses described in the next section that tease out the difference between the effects of assignment to treatment and receiving the treatment. But in an important sense, implementation issues are larger than what we have covered to this point. They may include inputs that the treatment requires, including client flow, treatment resources, provider time, and managerial support. They may include contextual issues such as local politics and social setting peculiarities that constrain how treatment is done. And they may include funding or insurance reimbursement rules for paying for treatment. Information about all of these matters is important for two reasons. One is to anticipate potential breakdowns in the intervention so that they can be monitored and prevented before they undermine the study; the other is to provide descriptions of the context of implementation that can be crucial to those who are considering using it in another context.

Two methods help accomplish these goals. One is to construct a process model of the treatment that portrays the intervention and its context, typically a figure complete with boxes for inputs, processes, and outputs, with causal arrows connecting them to portray the time flow of events that are supposed to occur. A number of authors have provided detailed instructions for how such models can be constructed, along with practical examples (Becker, 1992; Rossi, Freeman, & Lipsey, 1999; Sechrest et al., 1979; Weiss, 1998). The second method is good description of all these matters in study reports, the more detailed the better. In both cases, the use of these methods should occur from the start of the study, be maintained throughout its implementation, and be included in reports. In some cases, such as multisite experiments, for example, it may help to have an ongoing newsletter in which useful information about these issues is presented and in which individuals can exchange experiences regarding difficulties they have met in implementing their treatment or procedures they have adopted for alleviating these problems.

Treatment Implementation in Efficacy and Effectiveness Studies

Although much can be learned from *assessing* treatment implementation, *inducing* full treatment implementation is not always necessary or desirable. In particular, the internal validity of the inference that *assignment to condition caused outcome* does not require the treatment to be fully implemented. Researchers in public health recognize this fact in their distinction between tests of treatment efficacy versus effectiveness:

> Efficacy denotes the degree to which diagnostic and therapeutic procedures used in practice can be supported by scientific evidence of their usefulness under optimum conditions. Whether or not these procedures are applied adequately in practice, and whether they produce the intended results when so applied, are matters of effectiveness. (Starfield, 1977, p. 71)

In efficacy trials, treatments often are standardized, and full implementation is the goal, so the treatment is given every possible chance to show its effects. This procedure is particularly desirable when a treatment is first being studied because it would make little sense to pursue a treatment that does not perform satisfactorily under optimal circumstances. But in effectiveness trials, because researchers recognize that treatments are often administered in the real world with less than full standardization and implementation, inclusion criteria may be loosened and recipient compliance may be left to be variable, all because researchers want to know how a treatment will perform in such less-than-ideal circumstances. Indeed, haphazard standardization and implementation are so characteristic of many social interventions that stringent standardization would not well represent treatment in practice. A randomized trial in which treatment standardization and implementation are left to vary according to the contingencies of practice still yields an internally valid estimate of the effectiveness of that treatment-as-standardized-

and-implemented. However, the construct validity of the treatment characterization clearly depends on the nature of treatment implementation—indeed, the words *efficacy* and *effectiveness* are simply construct labels we choose to use to distinguish different kinds of conditions to which we assign people in experiments.

Analyses Taking Implementation into Account

When treatment implementation data are available, experimenters may analyze them in three ways:

- An intent-to-treat analysis.
- An analysis by amount of treatment actually received.
- By one of a variety of newly-developed analyses that try to combine some of the benefits of the first two options.

Intent-to-Treat Analysis

In an **intent-to-treat analysis**, participants are analyzed as if they received the treatment to which they were assigned (Begg, 2000; Lachin, 2000; Lavori, 1992; Lee, Ellenberg, Hirtz, & Nelson, 1991; Rubin, 1992a). This analysis preserves the benefits of random assignment for causal inference but yields an unbiased estimate only about the effects of being assigned to treatment, not of actually receiving treatment. The inference yielded by the intent-to-treat analysis is often of great policy interest because if a treatment is implemented widely as a matter of policy (say, by being mandated by law or funded by an insurance company), imperfect treatment implementation will occur. So the intent-to-treat analysis gives an idea of the likely effects of the treatment-as-implemented in policy. But the inference is not of universal interest. Moreover, the intent-to-treat analysis can yield biased results in the presence of nonrandom missing outcome data, requiring additional assumptions and analyses to yield valid effects (Frangakis & Rubin, 1999). Consequently, although researchers should conduct and report an intent-to-treat analysis, they should supplement it with other analyses.

Analysis by Amount of Treatment Received

If treatment implementation has been measured, the researcher can compare outcomes for those who received treatment with outcomes for those who did not. However, this comparison is quasi-experimental (compared with the intent-to-treat analysis) because participants were not assigned to receipt of treatment at random. For example, all things being equal, if outcome improves as the amount of treatment implemented increases within the treatment group (and within the comparison group, if relevant), the improvement constitutes only weak circumstantial evidence that the treatment caused the outcome. It is weak because peo-

ple may have self-selected into receiving greater levels of treatment based on, say, being more motivated than those who chose to receive lower levels of treatment. Treatment effects are then confounded with those unknown selection biases, just as they are in any quasi-experiment. Hence analyses by amount of treatment ought to be done in addition to, rather than instead of, a standard intent-to-treat analysis.

Instrumental Variable Analyses

The state of the art of these kinds of analyses is rapidly improving (West & Sagarin, 2000). In one of the most influential of these works, Angrist, Imbens, and Rubin (1996a) use random assignment as an **instrumental variable** (Foster & McLanahan, 1996) to obtain an unbiased estimate of the average causal effect for those who receive treatment.[1] They consider a randomized experiment with a binary outcome and a binary measure of whether participants in both conditions complied with treatment or not. They make five *strong* assumptions, three of which are straightforward and often plausible: (1) that one person's outcomes do not vary depending on the treatment someone else is assigned (e.g., if two friends were assigned to different conditions, one to get a flu vaccine and the other not, the probability of one friend getting the flu might decrease because the other was vaccinated); (2) that, by virtue of random assignment, the causal effects of assignment both on receipt and on outcome can be estimated using standard intent-to-treat analyses; and (3) that assignment to treatment has a nonzero effect on receipt of treatment. The remaining two assumptions can be problematic, but Angrist et al. (1996a) describe sensitivity analyses to explore the magnitude of bias that results from violations: (4) that random assignment (the instrumental variable) affects outcome only through its effects on receipt of treatment, and (5) that there are no "oppositional" participants who would always refuse treatment if assigned to it but take treatment if not assigned to it. Although both assumptions can be plausible, both have exceptions we discuss shortly.

Angrist et al. (1996a) illustrate the method and the assumptions with the Vietnam draft lottery, in which birth dates were assigned random numbers from 1 to 365, and those below a certain number were then subject to the draft (in effect, being randomly assigned to draft eligibility). However, not all those subject to the draft actually served in the military. Suppose the question of interest is whether serving in the military (not draft eligibility) increases mortality. The standard intent-to-treat analysis uses randomization to examine whether draft eligibility increases mortality, which

1. See comments on this method and example by Heckman (1996); Moffitt (1996); Robins and Greenland (1996); Rosenbaum (1996a); and Angrist, Imbens, and Rubin (1996b). Heitjan (1999) compares the method with others. Special cases of the method were presented more intuitively by Bloom (1984a) and Zelen (1979), though certain ethical issues can arise (Snowdon, Elbourne, & Garcia, 1999; Zelen, 1990) with some implementations of Zelen's randomized consent design in which participants are randomized to conditions before they have given informed consent (see also Braunholtz, 1999; Elbourne, Garcia, & Snowdon, 1999).

yields an unbiased estimate of effects but is not quite the question of interest. We can address the question of interest by comparing the mortality of those who served in the military with that of those who did not, but the comparison is then quasi-experimental, biased by the many unknown factors other than the draft that caused people to serve in the military (e.g., volunteering to maintain a family tradition, being cajoled by peers who enlisted). The Angrist et al. (1996a) method provides an unbiased instrumental variable estimate of the question of interest if the aforementioned assumptions are met. Clearly the first three assumptions are no less plausible than they are in any randomized experiment. As to the fourth, however, a potential draftee's knowledge that he was now eligible for the draft might cause him to stay in school to gain a deferment, which might improve mortality rates through education and income. The fifth assumption is that no one is so oppositional that if drafted he would refuse to serve but if not drafted he would volunteer. This is generally plausible, but one can imagine exceptions, such as the person whose family history would have encouraged him to volunteer for the military in the absence of being drafted but who objected to the government draft and so refused to serve in protest. If we know the prevalence of such violations of assumptions, sensitivity analyses can show the magnitude of expected biases. Using these analyses, Angrist et al. (1996a) showed that violations of the fourth assumption might significantly bias results in this example.

Variations on this method are rapidly appearing for use in studies with variable treatment intensity, such as drug dosage or hours of exam preparation; with multivalued instrumental variables; with providing bounds on estimates rather than point estimates; and with quasi-experiments and other observational studies (Angrist & Imbens, 1995; Balke & Pearl, 1997; Barnard, Du, Hill, & Rubin, 1998; Efron & Feldman, 1991; Fischer-Lapp & Goetghebeur, 1999; Goetghebeur & Molenberghs, 1996; Goetghebeur & Shapiro, 1996; Imbens & Rubin, 1997a, 1997b; Little & Yau, 1998; Ludwig, Duncan, & Hirschfield, 1998; Oakes et al., 1993; Robins, 1998; Rosenbaum, 1995a; Sommer & Zeger, 1991). The plausibility of assumptions may decrease in some of these applications. Developments on this topic are so rapid that readers are well-advised to search the literature before relying solely on the preceding references.

An issue is that the answer these methods yield may depend on the measure of implementation that is used. Heitjan (1999) shows that even a simple classification of a participant as a complier or noncomplier is fraught with subjectivity. Similarly, an implementation measure with low reliability might attenuate results, and a measure with low validity would presumably call the construct validity of the treatment into question. Further, these models currently use only one measure of implementation, but if one can develop several measures of delivery, receipt, and adherence, no single measure may best capture implementation (though it seems that adherence is the intended target in the Angrist et al. method). If several measures exist, the implementation analysis could be run repeatedly, once for each available implementation measure, and results inspected for more or less consistent results. A version of the method that could use several implementation measures simultaneously would be desirable.

POST-ASSIGNMENT ATTRITION

In this section, we start by defining the problem of post-assignment attrition and the difficulties it causes. Then we discuss how to, first, prevent attrition and, second, statistically analyze data when attrition has occurred.

Defining the Attrition Problem

Post-assignment attrition refers to any loss of response from participants that occurs after participants are randomly assigned to conditions. Such losses can range from an inadvertent failure to answer a single questionnaire item to the loss of all data on predictors and outcomes that occurs when a participant refuses any further participation. Post-assignment attrition should usually include cases in which, after assigning a participant to conditions, an experimenter deliberately drops that participant from the data. Such deliberate drops are a problem if they could have been caused by assignment. For example, researchers often drop participants for failing to meet an eligibility criterion. In some cases this is plausible, as when a study eligibility criterion was that all participants be female and a male was accidentally included. Treatment could not have caused gender, so dropping this person will cause no bias. In many other cases, however, this judgment is subject to considerable unreliability, as when a researcher decides to drop a participant as not really being bulimic or anorexic. The latter judgments are made with sufficient error to allow inadvertent biases based on experimenter or participant expectations. Even with such seemingly reliable rules as dropping those who move from the area, treatment assignments (e.g., to a less desirable control condition) can push a participant over the edge to decide to move, thereby making the move treatment-correlated. Given such ambiguities, it is rarely a good idea to deliberately drop participants after assignment to conditions.

All attrition lowers statistical power, and treatment-correlated attrition of participants from conditions threatens internal validity in a randomized experiment. Many of the benefits of random assignment occur because it creates equivalence of groups on expectations at pretest, an equivalence that is presumed to carry over to posttest. But when attrition is present, that equivalence may not carry over, particularly because attrition can rarely be assumed to be random with respect to outcome. In such cases, the nonrandom correlates of attrition may be confounded in unknown ways with treatment, compromising inference about whether treatment caused posttest outcomes.

Moderate to high attrition from treatment outcome studies has been reported in such widely diverse areas as smoking cessation (Klesges et al., 1988), alcohol treatment (Stout, Brown, Longabaugh, & Noel, 1996), substance abuse treatments (Hansen, Tobler, & Graham, 1990), psychotherapy (Epperson, Bushway, & Warman, 1983; Kazdin, Mazurick, & Bass, 1993; Weisz, Weiss, & Langmeyer, 1987), and early childhood education (Lazar & Darlington, 1982). In a meta-analysis of

85 longitudinal cohorts in adolescent substance abuse research, Hansen, Tobler, and Graham (1990) found that the average attrition rate was 18.6% at 3 months posttreatment and 32.5% at 3 years. Other reviews have found attrition rates greater than 40–50% in studies of substance abuse, homelessness, and child behavior problems (Ribisl et al., 1996).

Both lore and evidence suggest that attrition is often systematically biased rather than random (Bloom, 1990; Klesges et al., 1988; MacKenzie, Funderburk, Allen, & Stefan, 1987; Mennicke, Lent, & Burgoyne, 1988; Stout et al., 1996). For example, lore says that individuals who drop out of job training programs that guarantee an income during training are likely to be the more able, because they are more likely than others to find work that pays well; that individuals who drop out of parole programs will be those with the lowest chance of "rehabilitation"; or that individuals who drop out of an experimental group designed to evaluate the effectiveness of day care for the elderly will be the oldest and most infirm and those who are less gregarious. Data support such lore. Kazdin, Mazurick, and Bass (1993) found substantial differences between those who drop out of child psychotherapy and those who stay for treatment. MacKenzie, Funderburk, Allen, and Stefan (1987) found that alcoholics lost to follow-up differ from those who could be found and interviewed. Adolescents who drop out of a substance abuse study have different drug use patterns than those who stay (Brook, Cohen, & Gordon, 1983; Tebes, Snow, & Arthur, 1992). Examinees who leave questions unanswered on tests differ on a variety of personal characteristics from those who answer all questions (Grandy, 1987). Given such data, the burden of proof should be on the researcher to show that attrition is *not* treatment-correlated when it occurs and that the failure to find such a relationship is not due to low power.

By virtue of treatment-correlated attrition, many randomized experiments in practice become more similar to quasi-experiments. However, this does not mean that randomized experiments with attrition are no better than quasi-experiments. After all, these same attrition biases may exist in quasi-experiments, adding attrition biases to the selection biases already present in such designs. Meta-analytic evidence suggests that effect sizes from randomized experiments with attrition fall between those from randomized experiments with no attrition and quasi-experiments (Shadish & Ragsdale, 1996). It is likely, therefore, that initial randomization can often reduce the overall magnitude of posttest bias when compared with quasi-experimental approaches, even when attrition occurs.

Preventing Attrition

The most desirable solution to attrition is to prevent it from occurring. Preventing all attrition is rarely possible, but minimizing attrition is still important if the effects of attrition are cumulative. Not all attrition is preventable—for example, that due to natural disasters, death, and riots (Clarridge, Sheehy, & Hauser, 1977). And the costs of minimizing attrition could be used for other purposes, such as in-

creasing sample size, so balancing such expenditures is a consideration (Groves, 1989). However, much attrition can and should be prevented.

Attrition Caused by Treatment or by the Research Procedures

Treatment dropout is not unique to experiments. Psychotherapists refer to "premature termination" of therapy as a substantive problem, and in medicine, noncompliance with medication is common. Not surprisingly, then, some attrition from an experiment is due to features of treatment. Some women declined to be trained in elevator repair in the Rockefeller MFSP program because they perceived it to be a traditionally male job (Boruch et al., 1988). Many depressed outpatients refused randomization to a drug treatment condition that they found unattractive (Collins & Elkin, 1985). Counseling clients are less likely to drop out when their therapists are seen as expert, attractive, and trustworthy (Mennicke et al., 1988). Some of these features can be manipulated to decrease attrition, as with pretherapy training about the nature of treatment and the expectations a client should have, and by tailoring treatments to more closely match client expectations (such as offering brief therapy because many clients expect that). However, there are limits to how much manipulation can be done while still maintaining the original research question. One could not offer brief therapy if the question concerns an extended treatment. Sometimes the offending features of a treatment are inherent to it, as with the deliberate administration of nausea drugs in aversion therapy for sex offenders.

Other times attrition is caused by the research process. The demands of research exceed those normally expected by treatment recipients. An example is the tradeoff between the researcher's desire to measure many relevant constructs as accurately as possible and the respondent's desire to minimize the time spent answering questionnaires. Interpersonal conflict between research staff and participants can cause attrition, as can requests made of participants that exceed their resources, such as requiring mothers to leave their children at home without day care, research staff's failure to keep promises to participants such as providing feedback about how they performed on tests, participant fear of lack of confidentiality, and scheduling difficulties that the participant may have little incentive to resolve. Researchers can identify such problems in pilot studies before the main experiment begins, but even these fail to locate all such sources of attrition. After the fact, the researcher can identify these problems by **debriefing** participants about their study experience and by asking dropouts why they failed to return. These can be done by a staff member not involved in other aspects of the study, for those staff members might minimize reports of negative experiences that might reflect poorly on their work.

Retention and Tracking Strategies

Attrition is especially likely in studies of research populations that are mobile or afraid to make their locations known or that lack phones or employment, such as

the homeless, drug abusers, or abused spouses. Special efforts to retain and track these populations have been developed, and these efforts are equally applicable to any other populations. A thorough review with a rich trove of advice on retention and tracking strategies is provided by Ribisl et al. (1996; see also Boruch, 1997). They summarize their recommendations in eight parts (see Table 10.1 for detail): (1) gather complete location information at baseline from the participants, friends, or relatives and any available records or agencies that might know their whereabouts, including release of information forms giving permission to contact them; (2) establish formal and informal relationships with public and private agencies, such as driver's license bureaus, that might help find participants; (3) create a project identity with such means as a logo and identification badge; (4) emphasize the importance of tracking to project staff dedicated to that task and ensure that those staff are well-supported and compensated; (5) use the simplest and cheapest tracking methods first, saving more extensive methods for hard-to-find participants; (6) make research involvement convenient and rewarding for participants by such means as providing day care or lotteries for participation and by using alternative means of data collection, such as phone interviews if in-person interviews are not feasible; (7) expend the greatest amount of tracking effort at the initial follow-up periods, when most attrition occurs; and (8) customize tracking efforts to the individual participant's situation and to the study's circumstance. Such specifically tailored advice has been published on tracking and retention for prevention research (Biglan et al., 1991), childhood behavioral problems (Capaldi & Patterson, 1987), the homeless (Cohen et al., 1993), youth at high risk for AIDS (Gwadz & Rotheram-Borus, 1992), and smoking prevention research (Pirie et al., 1989).

Consider a longitudinal study that interviewed 141 abused women six times over 2 years (Sullivan, Rumptz, Campbell, Eby, & Davidson, 1996). Interviewers were trained in how to establish trust with these women, including explaining why the research was important and stressing the confidentiality of the women's identity, location, and response. Interviewers obtained names, addresses, and phone numbers of all contact persons who would know where the woman was living if she moved, including family, friends, neighbors, employers, coworkers, clergy, individuals in local organizations, and government agents, such as their social workers or the local social security office. Release of information forms giving permission to contact these people were obtained. The woman received a business card with information about how to contact the project (including a toll-free number) with the date of the next interview, the amount she would be paid for that interview, and a request to call the project if she moved. Finally, the project paid each woman an increasing amount of money for each interview as an incentive for her to make contact if she moved. Halfway between any two interviews, the project staff called the woman to remind her of the upcoming interview, and then a week before the interview would try to locate her and set up the interview. If phone calls, letters, and home visits did not locate the woman, then they called, visited, and sent letters to the alternative contacts and made trips into the woman's community to find people who might know her location.

TABLE 10.1 Comprehensive Listing of Retention and Tracking Techniques

Information collected from participant	
Demographics of participant	• First and last name, middle initial (or name) and all aliases, nicknames • Intentions to change name • Social Security number • Medicaid/Medicare number • Date of birth and place (city, town, state, hospital) • Home address(es), mailing address(es), and phone number(s) • Current and previous occupation, work address(es), and phone number(s) • Veteran status—if applicable, claim number and dates of service • Student status, name and address of school, school district • Driver's license number • Participant's moving plans in the next year
Demographics of relatives (parents, spouse, significant other, stepparents, siblings, children)[a]	Obtain the following information on at least two people: • Full names (maiden or birth names for women if married and changed name) • Addresses and phone numbers • Date of birth
Demographics of collaterals[a]	• Name, address, phone number of significant others/friends • Name, address, phone number of representative payee, state welfare worker, and religious contacts, if applicable • Name, phone number of landlord, if applicable
Demographics of professionals[a]	Identifying information for: • Community mental health caseworker or primary therapist • Department of Social Services caseworker • Parole/probation officer—note times and dates of incarceration • Medical doctor and/or clinics/hospitals utilized • Names of shelter workers or shelters the participant frequents

a. You must have a signed "Release of information" form for each of these individuals.

TABLE 10.1 Continued

Relatives, correspondents, and professionals (optional)	• Contact these people to check the accuracy of information • If the participant is institutionalized, verify contacts before his or her release and discuss with them any conflicting or inaccurate information
Retention and tracking procedures	
Participant	• Request that participant call project office if any tracking information changes or give him or her prepaid change of address cards • Call or write 2–3 weeks after initial contact (or midway between end of treatment and first follow-up), as the "trail is fresh," and contact again at least 2 weeks before follow-up data point • After locating the person, go over all current tracking information and add any new information (keep copies of all old information) • Schedule interview appointment for difficult-to-locate participants on the same day of contact or within 1 week, at a time most convenient for the participant • Give participant an interview card with the logo featuring: the name, address, and phone number of the study; a description of the incentive for completing an interview (if offered); time and place of next interview; and a reminder to call the office if any location information changes • Offer to help defray the costs that participants incur for getting to the interview (cab or bus fare, baby-sitting costs, etc.) • Send participants cards on their birthdays and any other appropriate occasions
Relatives/correspondents	• If you cannot locate the actual participant, call or write to relatives or correspondents • If the relative does not know the person's whereabouts, ask him or her if you can call again in a week or so, because people tend to resume contact eventually. Also send the collateral an interview card to forward to the participant if they see him or her • Contact other participants who were in treatment, jail, and so forth, at the same time as the participant (see McCoy & Nurco, 1991; Nurco, Robins, & O'Donnel, 1977, for a method to do this and still maintain confidentiality)

TABLE 10.1 Continued

Public records	• Telephone—check phone directory, call directory assistance, use criss-cross directory for participant and collaterals • Mail—contact post-office for change of address information for participants and collaterals, use U.S. Postal Service forwarding and record updates, certified mail, registered mail, stamped return envelope • Government records—check police and prison records, check parole and probation information—check names and addresses of visitors to jails or prisons. Contact marriage license bureaus; city, county, and state tax rolls; drivers' license bureaus; the Social Security Administration; the state welfare office; the FBI; population statistics. Contact State Departments of Public Health or Vital Statistics to inquire if the participant is deceased • Contact agencies: Alumni offices, local utility companies, high school records, professional organizations, treatment centers, credit agencies, psychiatric institutions, Veterans Administration hospital
Neighborhood environment	• Contact next-door neighbors when the participant has moved • Go to participant's home and workplace • Talk to landlord and neighbors and walk around the neighborhood and ask anyone "hanging out" • Go to corner restaurants, convenience stores
Preventing refusals	
	• Have same interviewer track participant over time to build rapport, or switch interviewers to see if participant likes the new one better • Provide immediate reinforcement for attending appointments • Provide snacks and beverages during interviews or treat participant if conducting interview at a restaurant or coffee shop • If participant has a history of missing interviews, send appointment reminder card or remind by telephone; mention incentives • Be nonjudgmental and open

TABLE 10.1 Continued

Potential dropouts	• Stress that all information will be held strictly confidential • Discuss incentives, if any, for participation • Remind the participant of the importance of the study and his or her participation • Have study director personally call or visit participant • Ask if you can call the participant back after a few days so he or she can think about the decision to drop out of the study • Do not coerce participant
Relatives, correspondents, and professionals	• From the beginning, have the participant inform contacts that researchers may contact them • Describe, mail, or show in-person the "Release of information" form that the participant has signed mentioning that person's name

Note: Not all of the information featured in this table typically needs to be collected, nor do all of the tracking procedures need to be attempted for all populations. The purpose of this list is to be as comprehensive as possible. The information in this table is from "Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations," by K. M. Ribisl, M. A. Walton, C. T. Mowbray, D. A. Luke, W. S. Davidson, and B. J. Bootsmiller, 1996, *Evaluation and Program Planning, 19,* pp. 1–25. Copyright 1996 by Elsevier. Adapted with permission.

With this intensive protocol, the researchers located about 95% of the women over 2 years. Phoning or going to the house of the woman more than once and phoning alternative contact persons accounted for 70–85% of successful contacts. The remaining strategies were necessary to get the overall success rate to 95%. If the project had relied solely on techniques that did not require them to leave the office, they would have lost 40% of their sample in the first 10 weeks after treatment—doubly crucial because this 40% of difficult-to-locate women had different characteristics than other women, and so would otherwise increase attrition bias greatly.

Preventing Treatment Attrition Versus Measurement Attrition

Measurement attrition (the topic of this section of the chapter) refers to a failure to complete outcome measurement, whether or not treatment is completed. **Treatment attrition** (part of the previous section on treatment implementation) refers to those research participants who do not continue in treatment, whether or not they continue taking the measurement protocol. In studies using archival

sources to measure outcomes such as arrest records, hospital mortality, or student grades, it is possible to obtain outcomes on everyone assigned to treatment, even those who did not complete treatment—such studies contain treatment attrition but no measurement attrition. Conversely, in brief studies that use captive participants (literally or figuratively), such as prisoners or elementary school students, it is possible to administer the treatment completely to every participant. But equipment failures, illness, objections to burdensome measurement, or careless responding may all lead to failure to obtain outcome measures—such studies contain measurement attrition but no treatment attrition.

The distinction is practically important for several reasons. First, measurement attrition prevents the inclusion of the participant in the analysis (except via missing data imputation methods), but treatment attrition does not preclude inclusion as long as the participant completed the measures. Second, many dropouts can be convinced to complete the measurement protocol even when they refuse to complete treatment. Third, if measurement attrition can be eliminated, the researcher can do a classic intent-to-treat analysis and, if a good implementation measure is available, can sometimes use the Angrist et al. (1996a) implementation analysis. So a good rule to follow is this: Prevent measurement attrition even when you cannot prevent treatment attrition.

Minimizing Time and Obstacles Between Randomization and Treatment

Attrition is lower when the time and obstacles between random assignment and treatment implementation are minimized. Bloom (1990), for example, reported results from three randomized experiments of programs in three different Texas cities to help employ displaced workers. In one site with participation rates of 87%, the intake process was quick, with little time and few obstacles between random assignment and receipt of treatment; in another site with more obstacles and time between those two steps, participation rates were only 60%. This leads naturally to the suggestion that assignment to treatment be delayed until the last possible moment.

For example, in a study to which one of us (Shadish) consulted, the aim was to examine the effects of different medications in preventing weight gain among participants who had just quit smoking. The study recruited smokers and asked all of them to quit over a 4-week period, during which they received identical quit-smoking interventions. The medication was administered in the 4th week, with both smoking cessation and medication continuing for several months. Originally, assignment was to occur at the start of Week 1, prior to smoking cessation. But a little thought shows that this may be too soon. The treatments to which participants were randomized were the different medications, not the program to help them quit smoking. The latter merely provided a pool of newly quit smokers on which to test the medications. In the final design, then, random assignment occurred at the start of Week 4 instead. Doing so minimized the attrition of participants who decided

they did not want to be in the study and of participants who could not quit smoking and thus could not be used in testing the main hypothesis. A very similar strategy uses a "running-in" procedure in which all participants are given one of the treatments, usually a standard treatment or one perceived as less desirable, for a period of time prior to randomization (Coyle, Boruch, & Turner, 1991). Many of those who would drop out of the study would have done so before the end of this run-in period. Those who remain are then randomized either to continue in the condition they began or to the innovative or more desirable treatment.

These methods to minimize attrition also increase selectivity of participants into the experiment, which may reduce generalizability. In the smoking cessation example, the causal inference about the effects of medication in reducing postcessation weight gain is more likely to generalize to smokers who were able to quit for more than 4 weeks and who were willing to be randomly assigned to different medications or a placebo. This excludes a substantial portion of the smoking population. However, given that the same experiment without a run-in period would still contain attrition, external validity would still be restricted by participant attrition. If the characteristics of preassignment dropouts and perhaps their reasons for dropping out are measured, valuable information about generalizability can be obtained, for example, about the kind of person who judges the treatment unacceptable. Methods outlined in Chapter 9 for getting better estimates of both treatment effects and generalizability can also be applied to this problem (Braver & Smith, 1996; Ellenberg, 1994; Marcus, 1997a).

Minimizing Treatment-Correlated Attrition

Differential attrition is more important than total attrition as a threat to internal validity. Differential attrition may be present even when equal percentages of treatment and comparison participants drop out, if they drop out for different reasons or have characteristics that relate to outcome. For example, in the New Jersey Negative Income Tax Experiment (Reicken et al., 1974), control group losses may have resulted from low economic motivation to cooperate, whereas the experimental losses may have resulted from unwillingness to accept charity. If so, equivalent attrition rates (7.9 versus 8.6) may mask differential selection factors that could produce group differences on outcome.[2]

Differential attrition occurs for many reasons. Differential vigilance by the research team during follow-up can keep more experimentals remaining in the study than controls. Similarly, no matter how late in the sequence randomization is deferred, some attrition may occur when treatments differ in desirability. For example, in the smoking and weight gain study described earlier, the medications were

2. Sometimes the dropout rate over conditions is an outcome of interest in its own right. In such a case, attrition rates are zero and can be analyzed with no attrition threat to internal validity.

administered to some participants through a chewing gum and to others using a skin patch. A disproportionate number of participants perceived the patch as more convenient and less unsightly and so refused their assignment to the chewing gum and dropped out of the study entirely. In studies using aggregate units such as schools, differential attrition can occur but can be harder to detect. For example, families may decide to remain in experimental school districts so that their children can receive an attractive treatment, but control families may leave more often; yet this might not be detected just by counting the number of schools in conditions (Reicken et al., 1974). When randomly assigning organized social units such as schools, it is particularly important to assign from among volunteers who agree to receive any treatment. This is because one is often dealing with so few units that the loss of any of them after randomization has serious consequences for the integrity of the design. Although this procedure does not guarantee that all units will accept the assignment—using this procedure, Cook, Hunt, and Murphy (2000) had one of 24 schools pull out after it learned it was a control—it reduces the problem considerably. But one must be willing to pay the price in terms of external validity.

One way to reduce treatment-correlated dropout is to use an informed-consent procedure in which the participant agrees to accept assignment to any experimental condition. However, this solution also reduces generalizability to similarly compliant ex-smokers; and it will not prevent the problem entirely because many participants will give consent in hopes of receiving the more desirable treatment, only to balk when they are assigned to the less desirable treatment. Here an option is a two-stage informed-consent procedure similar to that described by Reicken et al. (1974) in their commentary on the New Jersey Negative Income Tax Experiment. The first informed consent requests the participant's cooperation with measurement. Assignment to conditions is made from those who consent. The second informed consent requests agreement to the experimental treatments from those participants assigned to a treatment (but not necessarily from those assigned to the control condition unless an ethical constraint required it). Those who refused this consent are continued in the measurement protocol to which they already consented, reducing measurement attrition. However, those participants who agree to accept assignment to any condition are aware of all the treatment conditions by virtue of informed consent. So those who then receive less desirable treatments may envy recipients of the more desirable ones. This may lead to problems of compensatory rivalry or resentful demoralization. Sometimes this problem can be reduced by providing treatment-irrelevant incentives, such as more money or the chance to participate in a lottery for prizes, to those assigned to the less desirable condition.

Preventing Measurement Attrition

Reviews of the literature have suggested strategies for reducing measurement attrition depending on the kind of outcome measure used (Day, Dunt, & Day, 1995;

Lockhart, 1984; Yu & Cooper, 1983). They report that higher response rates occur with the use of personal or telephone (versus mail) surveys, the use of incentives to answer (either monetary or not, the more the better), providing prior notice of the questionnaire's arrival, using the foot-in-the-door method that gets the respondent to agree to a smaller task first and a larger one later, personalizing letters and other forms of contact, and using follow-up letters. Appealing to social values or flattering the respondents were not effective.

Some experiments take advantage of an established measurement framework that has been developed and maintained independently of the experiment, for example, court records or records of withholding tax. Even if a respondent drops out of the experiment, he or she can still be included in the measurement system. However, even with this approach some research participants will be missing from the record-keeping system, and the researcher is limited to using archival measures of unknown reliability and validity for answering the questions of interest.

A Flawed Approach—Replacing Dropouts

When participants drop out of an experiment, some researchers replace them with new participants who may even be randomly selected from the same pool of applicants as those originally assigned to conditions and who may also be randomly assigned to conditions (e.g., Snyder & Wills, 1989). Although this approach does keep the sample size at a specified level, which may be important for purposes of power, such replacement would only solve attrition as an internal validity threat if (1) both attrition and replacement are random, which is unlikely to be the case with attrition unless shown otherwise (and showing otherwise requires analyses that usually occur long after it would be feasible to add replacements), or (2) both former and replacement participants have the same latent characteristics, especially as pertains to outcome, which again is unlikely because we cannot know the latent characteristics of dropouts and so cannot match replacements to them. This problem is similar to the discussion in Chapter 4 of why matching does not substitute for random assignment (it only matches on observed but not latent characteristics), but in this case the selection bias is out of treatment rather than into it.

Analyses of Attrition

Ultimately, the goal of all analyses of attrition is to understand how much it threatens the validity of a conclusion about treatment effectiveness. There is no single best way to gain this understanding. So the researcher should usually do several analyses to shed light on the problem from different perspectives (see Table 10.2 for a summary). Unfortunately, reviews of the literature suggest that few studies undertake such analyses (Goodman & Blum, 1996).

TABLE 10.2 A Summary of Possible Attrition Analyses

Simple Descriptive Analyses

- Overall attrition rate
- Differential attrition rates for treatment and controls
- Whether those who completed the study differed from those who did not on important characteristics
- Whether those who completed the treatment group differed from those who did not on important characteristics and the same for the control group
- Whether those remaining in the treatment group differed from those remaining in the control on important characteristics

Identifying Different Patterns of Attrition

- Whether different groups have different patterns of attrition
- Whether different measures have different attrition patterns
- Whether certain subsets of respondents or sites have complete data that could be used to salvage some randomized comparisons

Accounting for Attrition When Estimating Effects

- Impute values for missing data
- Bracket the possible effects of attrition on effect estimates
- Compute effect estimates that are adjusted for attrition without using imputed data

Simple Descriptive Analyses

First, experimenters should provide simple descriptive data about attrition. A basic template for such analyses (Lazar & Darlington, 1982) reports (1) overall attrition rate, (2) differential attrition rates for treatment and controls, (3) whether those who completed the study differed from those who did not, (4) whether those who completed treatment differed from those who did not and the same for the control group, and (5) whether those remaining in treatment differed from those remaining in the control. For analyses (3) through (5), particularly important variables to analyze include reasons for dropping out, pretest scores on the measures that will subsequently be used as indicators of a treatment effect, and other pretest variables correlated with outcome. Pretests on the outcome measure provide the best available single estimates of the direction of bias for that outcome. After all, if the direction of the bias runs opposite to the empirical results of the study, attrition is much less of a threat to those results. For example, if those who scored best on the pretest dropped out differentially more often from treatment than from control, but posttest results suggest the treatment did better than the control despite this bias, it is less plausible to think that differential attrition is the cause of the observed effect. Moreover, the pretest difference of experimental and control units who remain frequently gives a helpful estimate of the magnitude of the pseudo-effect that would be expected at the posttest (Heinsman & Shadish, 1996;

Shadish & Ragsdale, 1996). This, in combination with the difficulty of maintaining randomly formed groups over time in complex field settings, is one reason why we recommend the use of pretests, even though they are not needed if randomization is successfully initiated and maintained without attrition.

Identifying Different Patterns of Attrition

A study may show several different patterns of attrition. Sometimes we define those patterns conceptually, as when we distinguish between those who could not be located, those who were located but would not complete the outcome variable, and those who died. Alternatively, we can search for patterns of missing data empirically. For example, BMDP's AM program explores data sets to try to discover missing data patterns[3]; and several stand-alone personal computer programs have extensive missing-data-analysis capabilities, including SOLAS (Statistical Solutions, 1998) and ESTIMATE (Marcantonio, 1998). Then the experimenter can explore whether different patterns of attrition have different predictors. MacKenzie, Funderburk, Allen, and Stefan (1987) found different predictors of attrition in alcohol treatment studies depending on how attrition was defined. Mennicke et al. (1988) found that early dropouts from counseling showed worse adjustment than later dropouts. Willett and Singer (1991) used survival analysis to differentially predict attrition that occurred at different times. Such definitional differences hold between studies, as well. Kazdin (1996) suggested that discrepancies among child psychotherapy researchers in their estimates of the rates and correlates of attrition was due partly to their use of different definitions of dropout.

Within the same study, different measures may have different attrition patterns. For instance, in the New Jersey Negative Income Tax Experiment (Reicken et al., 1974), Internal Revenue Service (IRS) records of labor force participation had a different attrition pattern than did personal interviews about labor force participation. Attrition from the IRS records occurred because some persons do not file tax returns and because employers may not file earnings data with the IRS in some cases. Some of this missing data might be found using periodic interviews. Attrition from the interviews was more strongly related to the financial value of the experimental treatment—more people refused to be interviewed in the less lucrative conditions. The latter point favors the IRS measures to the extent that any underrepresentation of persons in IRS records would probably affect experimentals and controls alike, which was not the case with interviews.

Sometimes treatment-correlated attrition is restricted to certain subgroups of respondents or sites. If examination of pretest data and attrition rates suggests that there are sites or subgroups where pretest comparability can reasonably be assumed, these deserve special attention in the data analysis. But although the search for randomly comparable subgroups or sites is worthwhile, it must be done

3. BMDP was purchased by SPSS in 1996 and is available through that company.

cautiously, because sample sizes will be reduced (making tests of comparability less powerful) and the possibility of capitalizing on chance will be increased if multiple tests are done without some correction for overall error rate. Nonetheless, careful disaggregation will sometimes help salvage some true experimental comparisons that have been compromised by attrition.

Accounting for Attrition When Estimating Effects

Here the goal is to estimate what the study would have shown had there been no attrition. Two general approaches are common: one that imputes values for the missing data points so that they can be included in the analyses and another that tries to estimate effects without imputing missing data.

Imputing Values for Missing Data. Programs such as SOLAS (Statistical Solutions, 1998) and ESTIMATE (Marcantonio, 1998) are specifically designed to do most of the imputations we cover here, and many other programs have some capacity to impute data, too (e.g., EQS; Bentler & Wu, 1995). Imputation methods vary from simple to complex, with the simpler methods yielding the least satisfactory imputations. For example, the simplest but least satisfactory approach is to substitute the sample mean on a variable for the missing data point. The implicit assumption underlying mean substitution is that data are missing at random,[4] which is usually implausible.

At the other extreme are methods that use maximum likelihood algorithms to estimate missing data, the best of which is multiple imputation (Jennrich & Schlucter, 1986; Rubin, 1987; Little & Rubin, 1987; Little & Schenker, 1995). The latter uses multiple regression to predict missing data points from a set of measured predictors, does this several times, and then adds random error to each imputed data point. The average of these multiple imputations is then used. However, these methods make strong distributional assumptions, and they generally assume that the mechanisms generating missing data are either known or ignorable, that is, unrelated to the actual value of the missing data, which is unlikely to be true in many settings. For example, people who drop out of a drug abuse study often do so because they return to drug use, which *is* the outcome variable often missing and needing imputation.

4. Little and Rubin (1987) distinguish between data missing completely at random (MCAR) versus data missing at random (MAR). Data are MCAR if data are missing for reasons that are completely unrelated to any information in the available data—randomly discarded data, for example. Data are MAR if they are missing for reasons related to the value of another observed variable that has no missing data, but specifically not because of the value on the missing variable itself. Because missing data on the outcome variable is often of interest, this latter requirement implies that data cannot be MAR if, say, drug addicts dropped out of a drug treatment study because they started using drugs again when drug use is the outcome of interest. If missing data are MCAR or MAR, they are ignorable for likelihood-based inferences; and any other kind of missing data are nonignorable. Ignorability implies that unbiased estimates of treatment effectiveness can be obtained with proper analyses (Little & Rubin, 1987). Ignorability is probably rare in most data sets.

In between are a host of other imputation options that are occasionally good choices for particular situations, such as the last value forward method for longitudinal data sets, hot deck imputation, and single imputation (Little & Schenker, 1995; Little & Yau, 1996; Speer, 1994; Speer & Swindle, 1982; Statistical Solutions, 1998). Most of these methods work well only when good predictors of missing data are available. If not, two strategies may help. One strategy is to gather additional data on a random sample of dropouts. In principle, this would give good data on which to impute the values of all missing participants (Graham & Donaldson, 1993) and would yield much other valuable information pertaining to attrition. A key problem with this approach is whether small samples of missing participants would yield sufficiently small confidence intervals. If not, resources may be required to examine virtually all dropouts, which is prohibitive. However, preliminary work suggests that the samples might not have to be too large (Graham & Donaldson, 1993; Hansen & Hurwitz, 1996).

The other strategy is to try to bracket the effect within a range of plausible values. When using model-based predictions such as hot decking or multiple imputation, one can vary the imputation model to see how different assumptions about predictors change the imputed values and the treatment effect estimate (Little & Yau, 1996). When outcomes are dichotomous (success-failure, pass-fail, married-divorced, diseased–disease-free) or can reasonably be dichotomized for purposes of exploring the effects of missing data, an option is to explore the distribution of all possible outcomes that could have been observed under different assumptions about attrition (Delucchi, 1994; Shadish, Hu, Glaser, Kownacki, & Wong, 1998; Yeaton, Wortman, & Langberg, 1983). Shih and Quan (1997) present a similar method for bracketing the effects of attrition when outcome data are continuous; but their method depends on having covariates capable of equating treatment and control completers, something that cannot always be assumed.

Estimating Effects in Data Sets with Attrition. Though no clear winner has emerged here, a number of approaches are being investigated to estimate effects in data sets with attrition, and researchers should consider trying several of these. Allison (1987) and Muthen, Kaplan, and Hollis (1987) use multigroup structural equation models to estimate effects in the presence of missing data. The idea is to identify a small number of groups with different missing data patterns (perhaps using the methods discussed previously) and then analyze the same model for all groups in a multigroup analysis to see whether parameter estimates are stable across all groups despite the missing data. However, if sample size per group is small, results may not be stable; and if the number of groups is too large, the technique may be too computationally intensive. If the data are missing at random, the approach yields accurate inferences about parameters. Even when this condition is not met, as will generally be the case, the technique yields useful information about whether groups with different missing data patterns yield the same parameter estimates. Several structural equation modeling programs have examples of how to implement this approach (e.g., Arbuckle, 1997; Bentler, 1995).

In longitudinal designs with large numbers of repeated measures, several analytic techniques can be tried without imputing missing data. Kraemer and Thiemann (1989; see also S. Maxwell, 1998) propose estimating the slope (rather than a change score or an endpoint score) as the outcome in a design having many repeated measures, though the power of this method may be poor (Delucchi & Bostrom, 1999). The slope can be estimated from as few as two time points, so some missing data can be tolerated and the slope can still be estimated accurately. Similarly, growth curve analysis techniques do not require observations at all time points on all research participants or even that the observations be taken at fixed time points (e.g., Bryk & Raudenbush, 1992; Rogosa, 1988). However, in both these techniques nonrandom missing data points can bias results.

Economists have proposed methods to estimate treatment effects in the face of attrition that model the dropout process itself (Leaf, DiGiuseppe, Mass, & Alington, 1993; Welch, Frank, & Costello, 1983). Doing this well requires the researcher to think about and measure why people drop out of conditions. However, these models make strong assumptions about normal distributions and about finding variables that affect attrition but not outcome. When these assumptions are not met, these models may produce substantially worse estimates than no correction at all (Hartman, 1991; Stolzenberg & Relles, 1990). Some economists have proposed semi-parametric methods making fewer assumptions (e.g., Scharfstein, Rotnitzky, & Robins, 1999), though questions still arise about bias in these methods (Little & Rubin, 1999). Little (1995) presents a method for repeated measures designs that combines these econometric models with missing data imputation.

As with imputation, it is often good practice to conduct a variety of analyses under different assumptions about the nature of attrition, finishing up with a range of estimates of effect (e.g., Scharfstein et al., 1999). In some situations, all of the estimates will be in essential agreement and final inference will be easy. But in others, the estimates will bracket a range of contradictory outcomes and will thus serve as a warning that attrition artifacts may be masquerading as treatment effects.

Economists have also suggested some relatively simple diagnostic tests to examine whether attrition may be a problem in a study. Verbeek and Nijman (1992) suggest comparing fixed and random effects estimates of model parameters and comparing results on respondents with complete data to those for the sample as a whole (including dropouts). In both comparisons, the resulting estimates should not differ significantly if attrition is not biasing results. They also suggest certain variations to the correction factor used in the econometric models that should also be nonsignificant if attrition is not a problem. Foster and Bickman (1996) illustrate the use of these methods using data from a large quasi-experiment. If these tests indicate that attrition is a problem, the researcher may still be left with a puzzle as to what the correct effect size estimate really is.

It should be clear from this discussion that a wide array of options now exists for constructively analyzing experiments with missing data. When both total and differential attrition are low (e.g., less than 10%) *and* the effect size is high, these analyses will rarely change the qualitative conclusion about whether treatment

works compared with analyses that do not take attrition into account. Under most other circumstances, however, it is possible for the qualitative conclusion to change completely (e.g., Shadish et al., 1998), in which case careful attention to the results of multiple attrition analyses is required to understand the implications of attrition.

DISCUSSION

The material that we have covered to this point in the book is in some sense just an extension of the tradition of field experimentation begun by Campbell (1957) and reflected in Campbell and Stanley (1963) and Cook and Campbell (1979). As with those prior works, the primary focus has been on how to use experimental design to improve internal validity, though we have often tried to show how these design choices can also affect all the other validity types at the same time. In the next three chapters of the book, however, we move considerably beyond this traditional focus by providing a conceptual analysis and practical methods for improving construct and external validity. We do so in a way that remains consistent with the history of Campbell's ideas and preferences, especially in focusing on how randomized and quasi-experimental designs can still be used to support such generalizations.

11

Generalized Causal Inference: A Grounded Theory

Gen·er·al·ize (jĕn´ər-ə-līz´): v. **gen·er·al·ized, gen·er·al·iz·ing, gen·er·al·iz·es.** v. tr. 1. a. To reduce to a general form, class, or law. b. To render indefinite or unspecific. 2. a. To infer from many particulars. b. To draw inferences or a general conclusion from. 3. a. To make generally or universally applicable. b. To popularize.

Ground·ed (ground´ĕd): v. tr. 1. To place on or cause to touch the ground. 2. To provide a basis for (a theory, for example); justify. 3. To supply with basic information; instruct in fundamentals.

THE STRENGTH of the randomized experiment is its capacity to facilitate the causal inference that a change in A caused a change in the probability that B would occur. However, one of the oldest criticisms of the randomized experiment is that it is so locally defined that this clear causal inference comes at the cost of one's ability to generalize the causal connection found. Fisher (1935) himself criticized the randomized experiment on this point, and it is evident in the Campbell and Stanley (1963) claim that internal and external validity can be negatively related. Campbell's (1986) relabeling of internal validity as local molar causal validity highlights the same point—the embeddedness of experimental results in a particular local context seems to provide little basis for generalizing results beyond that context.

As we saw in Chapters 1 and 3, generalizing a causal inference involves generalizing about four entities—treatments, outcomes, units (usually persons), and settings. We also saw that we can make two different kinds of generalizations about each of these four entities: (1) generalizations about the constructs associated with the particular persons, settings, treatments, and outcomes used in the study (construct validity) and (2) generalizations about the extent to which the causal relationship holds over variation in persons, settings, treatment, and measurement variables (external validity). This chapter presents a grounded theory of