8

Randomized Experiments: Rationale, Designs, and Conditions Conducive to Doing Them

Ran·dom (răn′dəm): [From at random, by chance, at great speed, from Middle English *randon,* speed, violence, from Old French from *randir,* to *run,* of Germanic origin.] adj. 1. Having no specific pattern, purpose, or objective: *random movements; a random choice.* See Synonyms at chance. 2. Statistics. Of or relating to the same or equal chances or probability of occurrence for each member of a group.

DOES EARLY preschool intervention with disadvantaged children improve their later life? The Perry Preschool Program experiment, begun in 1962, studied this question with 128 low-income African-American children who were randomly assigned to receive either a structured preschool program or no treatment. Ninety-five percent of participants were followed to age 27, and it was found that treatment group participants did significantly better than controls in employment, high school graduation, arrest records, home ownership, welfare receipt, and earnings (Schweinhart, Barnes, & Weikart, 1993), although early IQ and academic aptitude gains were not maintained into early adulthood. Along with other experimental evidence on the effects of preschool interventions (e.g., Olds et al., 1997; Olds et al., 1998), these results helped marshal continued political support and funding for programs such as Head Start in the United States. In this chapter, we present the basic logic and design of randomized experiments such as this one, and we analyze the conditions under which it is less difficult to implement them outside the laboratory.

In the natural sciences, scientists introduce an intervention under circumstances in which no other variables are confounded with its introduction. They then look to see how things change—for instance, whether an increase in heat af-

fects the pressure of a gas. To study this, a scientist might place the gas in a fixed enclosure, measure the pressure, heat the gas, and then measure the pressure again to see if it has changed. The gas is placed in the enclosure to isolate it from anything else that would affect the pressure inside. But even in this simple example, the intervention is still a molar treatment package that is difficult to explicate fully. The enclosure is made of a certain material, the heat comes from a certain kind of burner, the humidity is at a certain level, and so forth. Full control and full isolation of the "intended" treatment are difficult, even in the natural sciences.

In much social research, more formidable control problems make successful experimentation even more difficult. For example, it is impossible to isolate a person from her family in order to "remove" the influences of family. Even in agricultural tests of a new seed, the plot on which those seeds are planted cannot be isolated from its drainage or soil. So many scientists rely on an approach to experimental control that is different from physical isolation—random assignment. The randomized experiment has its primary systematic roots in the agricultural work of statistician R. A. Fisher (1925, 1926, 1935; see Cowles, 1989, for a history of Fisher's work). Randomization was sometimes used earlier (e.g., Dehue, 2000; Gosnell, 1927; Hacking, 1988; Hrobjartsson, Gotzche, & Gluud, 1998; McCall, 1923; Peirce & Jastrow, 1884; Richet, 1884; Stigler, 1986). But Fisher explicated the statistical rationale and analyses that tie causal inference to the physical randomization of units to conditions in an experiment (Fisher, 1999).

THE THEORY OF RANDOM ASSIGNMENT

Random assignment reduces the plausibility of alternative explanations for observed effects. In this, it is like other design features such as pretests, cohorts, or nonequivalent dependent variables. But random assignment is distinguished from those features by one very special characteristic shared only with the regression discontinuity design: it can yield unbiased estimates of the average treatment effect (Rosenbaum, 1995a).[1] Moreover, it does this with greater efficiency than the

1. Three observations about the phrase "unbiased estimates of average treatment effects" are worth noting. First, some statisticians would prefer to describe the advantage of randomization as yielding a consistent estimator (one that converges on its population parameter as sample size increases), especially because we never have the infinite number of samples suggested by the theory of expectations discussed shortly in this chapter. We use the term *unbiased* in this book primarily because it will be more intuitively understood by nonstatistical readers and because it fits better with the qualitative logic of bias control that undergirds our validity typology. Second, in a random sampling model, sample means are always unbiased estimates of population means, so differences between sample means are always unbiased estimates of differences between population means. The latter estimates can be obtained without using random assignment. But such estimates are not the same as unbiased estimates of treatment effects. It is the latter that random assignment facilitates; hence its ability to facilitate the causal inference that we refer to with the shorthand phrase "unbiased estimates of treatment effects." Third, the phrase correctly refers to the average effect over units in the study, as distinguished from the effects on each unit in the study, which is not tested in a randomized experiment.

regression discontinuity design in a greater diversity of applications. Because unbiased and efficient causal inference is a goal of experimental research, it is crucial that researchers understand what random assignment is and how it works.

What Is Random Assignment?

Random assignment is achieved by any procedure that assigns units to conditions based only on chance, in which each unit has a nonzero probability of being assigned to a condition. A well-known random assignment procedure is a coin toss. On any given toss, a fair coin has a known (50%) chance of coming up heads. In an experiment with two conditions, if heads comes up for any unit, then that unit goes into the treatment condition; but if tails comes up, then it becomes a control unit. Another random assignment procedure is the roll of a fair die that has the numbers 1 through 6 on its sides. Any number from 1 to 6 has a known (1/6) chance of coming up, but exactly which number comes up on a roll is entirely up to chance. Later we recommend more formal randomization procedures, such as the use of tables of random numbers. But coin tosses and dice rolls are well-known and intuitively plausible introductions to randomization.

Random *assignment* is not random *sampling*. We draw random samples of units from a population by chance in public opinion polls when we ask random samples of people about their opinions. Random sampling ensures that answers from the sample approximate what we would have gotten had we asked everyone in the population. Random assignment, by contrast, facilitates causal inference by making samples randomly similar to *each other*, whereas random sampling makes a sample similar to *a population*. The two procedures share the idea of "randomness," but the purposes of this randomness are quite different.

Why Randomization Works

The literature contains several complementary statistical and conceptual explanations for why and how random assignment facilitates causal inference:

- It ensures that alternative causes are not confounded with a unit's treatment condition.
- It reduces the plausibility of threats to validity by distributing them randomly over conditions.
- It equates groups on the expected value of all variables at pretest, measured or not.
- It allows the researcher to know and model the selection process correctly.
- It allows computation of a valid estimate of error variance that is also orthogonal to treatment.

These seemingly different explanations are actually closely related. None of them by itself completely captures what random assignment does, but each sheds light on part of the explanation.

Random Assignment and Threats to Internal Validity

If treatment groups could be equated before treatment, and if they were different after treatment, then pretest selection differences could not be a cause of observed posttest differences. Given equal groups at pretest, the control group posttest serves as a source of counterfactual inference for the treatment group posttest, within limits we elaborate later. Note that the logic of causal inference is at work here. The temporal structure of the experiment ensures that cause precedes effect. Whether cause covaries with effect is easily checked in the data within known probabilities. The remaining task is to show that most alternative explanations of the cause-effect relationship are implausible. The randomized experiment does so by distributing these threats randomly over conditions. So treatment units will tend to have the same average characteristics as those not receiving treatment. The only systematic difference between conditions is treatment.

For example, consider a study of the effects of psychotherapy on stress. Stress has many alternative causes, such as illness, marital conflict, job loss, arguments with colleagues, and the death of a parent. Even positive events, such as getting a new job or getting married, cause stress. The experimenter must ensure that none of these alternative causes is confounded with receiving psychotherapy, because then one could not tell whether it was psychotherapy or one of the confounds that caused any differences at posttest. Random assignment ensures that every client who receives psychotherapy is equally likely as every client in the control group to have experienced, say, a new job or a recent divorce. Random assignment does not prevent these alternative causes (e.g., divorce) from occurring; nor does it isolate the units from the occurrence of such events. People in a randomized experiment still get divorces and new jobs. Random assignment simply ensures that such events are no more likely to happen to treatment clients than to control clients. As a result, if psychotherapy clients report less stress than control clients at posttest, the cause of that difference is unlikely to be that one group had more new jobs or divorces, because such stressors are equally likely in both groups. The only systematic difference left to explain the result is the treatment.

The only internal validity threat that randomization prevents from occurring is selection bias, which it rules out by definition, because selection bias implies that a systematically biased method was used for selecting units into groups but chance can have no such systematic bias. As for the other internal validity threats, randomization does not prevent units from maturing or regressing; nor does it prevent events other than treatment from occurring after the study begins (i.e., history). Pretests can still cause a testing effect, and changes in instrumentation can still occur. Random assignment simply reduces the likelihood that these threats are confounded with treatment.

Equating Groups on Expectation

In statistics, the preceeding explanation is often summarized by saying that random assignment equates groups on expectation at pretest. What does this mean? First, it does not mean that random assignment equates units on *observed* pretest scores. Howard, Krause, and Orlinsky (1986) remind us that when a deck of 52 playing cards is well shuffled, some players will still be dealt a better set of cards than others. This is called the luck of the draw by card players (and sampling error by statisticians). In card games, we do not expect every player to receive equally good cards for each hand, but we do expect the cards to be equal in the long run over many hands. All this is true of the randomized experiment. In any given experiment, observed pretest means will differ due to luck of the draw when some conditions are dealt a better set of participants than others. But we can expect that participants will be equal over conditions in the long run over many randomized experiments.

Technically, then, random assignment equates groups on the *expectation* of group means at pretest—that is, on the mean of the distribution of all possible sample means resulting from all possible random assignments of units to conditions. Imagine that a researcher randomly assigned units to treatment and control conditions in one study and then computed a sample mean on some variable for both conditions. These two means will almost certainly be different due to sampling error—the luck of the draw. But suppose the researcher repeated this process a second time, recorded the result, and continued to do this a very large number of times. At the end, the researcher would have a distribution of means for the treatment group over the samplings achieved and also one for the control group. Some of the treatment group means would be larger than others; the same would be true for the control group. But the average of all the means for the treatment group would be the same as the average of all the means for the control group. Thus the expectation to which the definition of random assignment is linked involves the mean of all possible means, not the particular means achieved in a single study.

When random differences do exist in observed pretest means, those differences will influence the results of the study. For example, if clients assigned to psychotherapy start off more depressed than those assigned to the control group despite random assignment, and if psychotherapy reduces depression, posttest depression scores might still be equal in both treatment and control groups because of the pretest group differences. Posttest differences between treatment and control groups then might suggest no treatment effect when treatment did, in fact, have an effect that was masked by sampling error in random assignment. More generally, the results of any individual randomized experiment will differ somewhat from the population effects by virtue of these chance pretest differences. Thus summaries of results from multiple randomized experiments on the same topic (as in psychotherapy meta-analysis) can yield more accurate estimates of treatment effects than any individual study. Even so, we still say that the estimate

from an individual study is unbiased. Unbiased simply means that any differences between the observed effects and the population effect are the result of chance; it does not mean that the results of the individual study are identical to the "true" population effect.

The preceding explanation uses pretest means to illustrate how randomization works. However, this is merely a teaching device, and the use of actual measured pretests is irrelevant to the logic. Randomization equates groups on expectations of *every variable before treatment, whether observed or not.* In practice, of course, pretests are very useful because they allow better diagnosis of and adjustment for attrition, they facilitate the use of statistical techniques that increase statistical power, and they can be used to examine whether treatment is equally effective at different levels of the pretest.

Additional Statistical Explanations of How Random Assignment Works

Randomization ensures that confounding variables are unlikely to be *correlated* with the treatment condition a unit receives. That is, whether a coin toss comes up heads or tails is unrelated to whether you are divorced, nervous, old, male, or anything else. Consequently, we can predict that the pretest correlation between treatment assignment and potential confounding variables should not be significantly different from zero.

This zero correlation is very useful statistically. To understand this requires a digression into how to estimate treatment effects in linear models. Let us distinguish between the *study* and the *analysis* of the study. In a *study* of the effects of psychotherapy, stress is the dependent variable (Y_i), psychotherapy is the independent variable (Z_i), and potential confounds are contained in an error term (e_i). In the *analysis* of that study, the effects of treatment are estimated from the linear model:

$$Y_i = \mu + \hat{\beta}Z_i + e_i \qquad (8.1)$$

where μ is a constant, $\hat{\beta}$ is a regression coefficient, and the subscript i ranges from 1 to n, where n is the number of units in the study. Thus Y_i is the score of the ith unit on a measure of stress, Z_i is scored as 1 if the unit is in psychotherapy and 0 if not, and e_i consists of all potential confounding variables. In the analysis, if $\hat{\beta}$ is significantly different from zero, then psychotherapy had a significant effect on stress, and $\hat{\beta}$ measures the magnitude and direction of that effect.

For all this to work properly, however, the model that is specified in the analysis must match the reality of the study. Failure to achieve this match is called **specification error**—an incorrect specification of the model presumed to give rise to the data. Specifically, the statistical techniques used to estimate models such as equation (8.1) choose values of $\hat{\beta}$ so that correlations between the resulting errors and the predictor variables are zero (Reichardt & Gollob, 1986). The statistics do

this whether or not that correlation really was zero in the study. Fortunately, random assignment assures that the correlation in the study will be zero for reasons outlined in the previous section; so the study matches the analysis. However, in nonrandomized studies, many confounds are probably correlated with receipt of treatment, but the computer program still chooses $\hat{\beta}$ so that the error is minimally correlated with the predictors in the data analysis, yielding a mismatch between the study and the analysis. The result is an incorrect estimate of treatment effects.[2]

A related way of thinking about the benefit of randomization is that it provides a valid estimate of error variance (e.g., Keppel, 1991; R. Kirk, 1982). Two possible causes of total variation in outcome (i.e., of how much people differ from each other in stress levels) exist—variation caused by treatment conditions (e.g., whether the person received psychotherapy) and variation caused by other factors (e.g., all the other causes of stress). Random assignment allows us to separate out these two sources of variability. Error variation is estimated as the amount of variation among units within each condition. For example, for those clients who were assigned to psychotherapy, variation in whether or not they received psychotherapy cannot contribute to their different stress levels because there was no such variation—they all got psychotherapy. So any variance in outcome among people randomly assigned to psychotherapy must be caused only by confounds. The average of each of these computed error terms from within each condition serves as our best estimate of error. This error term is the baseline against which we see if differences *between* treatment conditions exceed the differences that normally occur among units as a function of all the other causes of the outcome.

Summary

Random assignment facilitates causal inference in many ways—by equating groups before treatment begins, by making alternative explanations implausible, by creating error terms that are uncorrelated with treatment variables, and by allowing valid estimates of error terms. These are interrelated explanations. For example, groups that are equated before treatment begins allow fewer alternative explanations if differences later emerge, and uncorrelated errors are necessary to estimate the size of the error term. But randomization is not the only way to accomplish these things. Alternative explanations can sometimes be made implausible through logical means, as is typically the aim with quasi-experimentation; and uncorrelated errors can be created with other forms of controlled assignment to conditions, as with the regression discontinuity design. But randomization is the only design feature that accomplishes all of these goals at once, and it does so more reliably and with better known properties than any alternatives.

2. One way to think about the selection bias models in Chapter 5 is that they try to make the error terms orthogonal to the predictors in a statistically acceptable way, but this is hard, so they often fail; and one way to think about the regression discontinuity design is that it is able to make this correlation zero for reasons outlined in the Appendix to Chapter 7.

Random Assignment and Units of Randomization

We have frequently used the word "unit" to describe whatever or whoever is assigned to experimental conditions. A unit is simply "an opportunity to apply or withhold the treatment" (Rosenbaum, 1995a, p. 17).

Kinds of Units

In much field experimentation, the units being assigned to conditions are people—clients in psychotherapy, patients in cancer trials, or students in educational studies. But units can be other kinds of entities (Boruch & Foley, 2000). R. A. Fisher (1925) assigned plots of land randomly to different levels of fertilizer or different strains of seed. In psychological and medical research, animals are often randomly assigned to conditions. Researchers in the New Jersey Negative Income Tax experiment (Rees, 1974) randomly assigned families to conditions. Gosnell (1927) randomly assigned neighborhoods to conditions. Edgington (1987) discussed single-participant designs in which treatment times were randomly assigned. Schools have been randomly assigned (Cook et al., 1998; Cook, Hunt & Murphy, 2000). Nor is randomization useful just in the social sciences. Wilson (1952) describes a study in which the steel plates used in gauges were randomized prior to testing different explosives, so that variations in the strength of the plates would not be systematically associated with any one explosive. The possibilities are endless.

Higher Order Units

Units such as families, work sites, classrooms, psychotherapy groups, hospital wards, neighborhoods, or communities are aggregates of individual units such as family members, employees, students, clients, patients, neighbors, or residents. Studies of the effects of treatments on such higher order units are common, and a literature has developed specific to experiments on higher order units (e.g., Donner & Klar, 2000; Gail, Mark, Carroll, Green, & Pee, 1996; Moerbeek, van Breukelen, & Berger, 2000; Murray, 1998; Sorensen, Emmons, Hunt, & Johnston, 1998). For example, the National Home Health Agency Prospective Payment Demonstration experiment assigned 142 home health agencies to different Medicare payment options to see how use of care was affected (Goldberg, 1997); the San Diego Nursing Home Incentive Reimbursement Experiment assigned 36 nursing homes to different Medicare reimbursement options (Jones & Meiners, 1986); the Tennessee Class Size Experiment randomly assigned 347 classes to large or small numbers of students (Finn & Achilles, 1990); and Kelly et al. (1997) randomly assigned eight cities to two conditions to study an HIV prevention intervention. The higher order unit need not be a naturally occurring entity such as a work site or a neighborhood. The researcher can create the higher order unit solely for the research, as in the case of a stop-smoking program that is administered in small groups so that participants can benefit from mutual support. Nor is

it necessary that the individual units know or interact with each other. For instance, when physicians' practices are randomized to conditions, the physician's practice is a higher order unit even though the majority of the physician's patients will never meet. Finally, sometimes a treatment cannot be restricted to particular individuals by its very nature. For example, when a radio-based driving safety campaign is broadcast over a listening area, the entire area receives treatment, even if only some individual drivers are formally included in the research (Reicken et al., 1974).

There are often good practical and scientific reasons to use aggregate units. In a factory experiment, it may not be practical to isolate each worker and give him or her a unique treatment, for resentful demoralization or diffusion of treatment might result. Similarly, in the first evaluation of "Plaza Sesamo," Diaz-Guerro and Holtzmann (1974) randomly assigned some individual children in Mexican day care centers to watch "Plaza Sesamo" in small groups. They were in a special room with two adult monitors who focused attention on the show. At the same time, other children watched cartoons in larger groups in the regular room with no special monitors. Because treating classmates in these different ways may have led to a focused inequity, it would have been desirable if the experimenters' resources had permitted them to assign entire classes to treatments.

The research question also determines at which level of aggregation units should be randomized. If effects on individuals are at issue, the individual should be the unit, if possible. But if school or neighborhood phenomena are involved or if the intervention is necessarily performed on an aggregate, then the unit of randomization should not be at a lower level of aggregation.[3] Thus, if one is investigating whether frequent police car patrols deter crime in a neighborhood, different amounts of patrolling should be assigned to neighborhoods and not, say, to blocks within neighborhoods.

In aggregate units, the individual units within aggregates may no longer be independent of each other because they are exposed to common influences besides treatment. For example, students within classrooms talk to each other, have the same teacher, and may all receive treatment at the same time of day. These dependencies lead to what used to be called the **unit of analysis** problem (Koepke & Flay, 1989) but what is more recently discussed as multilevel models or hierarchical linear models. Because this book focuses on design rather than analysis, we do not treat the analytic issues in detail (Feldman, McKinlay, & Niknian, 1996; Gail et al., 1996; Green et al., 1995; Murray, 1998; Murray et al., 1994; Murray, Moskowitz, & Dent, 1996). But from a design perspective, using higher order units raises several issues.

3. The nesting of participants in higher order units can still pose problems even when individuals are assigned to treatment. For example, if individual cancer patients who each have multiple tumors are randomly assigned to treatment but treatment is administered separately to each tumor and tumor response is observed separately for each tumor, those responses are not independent (Sargent, Sloan, & Cha, 1999).

Studies that use higher order units frequently have fewer such units available to randomize. Consider the limiting case in which students in one classroom are given the treatment and those in a second classroom serve as controls. Treatment conditions are then totally confounded with classrooms, making it impossible to tell if performance differences at posttest are due to differences in treatment or in classroom characteristics, such as the charisma of the teacher, the mix of students, or the physical conditions of the class. When more than one, but still few, higher order units are assigned to conditions, randomization may result in very different means, variances, and sample sizes across conditions. Such cases are surprisingly common in the literature (e.g., Simpson, Klar, & Donner, 1995); but they incur substantial problems for internal and statistical conclusion validity (Varnell, Murray, & Baker, in press). Such problems occur most often with studies of schools and communities, because it is expensive to add new sites. Random assignment of higher order units from within blocks or strata can reduce such problems. For example, McKay, Sinisterra, McKay, Gomez, and Lloreda (1978) studied the effects of five levels of a program of nutrition, health care, and education on the cognitive ability of chronically undernourished children in Cali, Colombia. They divided Cali into 20 relatively homogeneous neighborhood sectors. Then they rank-ordered sectors on a standardized combination of pretreatment screening scores and randomly assigned those sectors to the five conditions from blocks of five. The Kansas City Preventive Patrol experiment followed a similar procedure in its study of whether the visible presence of police patrols deterred crime (Kelling, Pate, Dieckman, & Brown, 1976). The researchers placed 15 patrol districts into blocks of three that were homogenous on demographic characteristics; they then randomly assigned districts from these blocks into the three experimental conditions.

Planning proper sample size and analysis of designs with higher order units is more complex than usual because individual units are not independent within aggregate units (Bock, 1989; Bryk & Raudenbush, 1992; Bryk, Raudenbush, & Congdon, 1996; H. Goldstein, 1987; Raudenbush, 1997; Snijders & Bosker, 1999). Given the same number of individual units, power is almost always lower in designs with higher order units than in those with individual units; and special power analyses must be used.[4] Moreover, power is improved more by increasing the number of aggregate units (e.g., adding more classrooms) than by increasing the number of individuals within units (e.g., adding more students within classrooms). Indeed, at a certain point the latter can rapidly become wasteful of resources without improving power at all, depending on the size of the dependencies within cluster (as measured by the intraclass correlation).

4. See Donner (1992); Donner and Klar (1994); Feldman et al. (1996); Gail, Byar, Pechacek, and Corle (1992); Gail et al. (1996); Hannan and Murray (1996); Koepsell et al. (1991); Murray (1998); Murray and Hannan (1990); Murray, Hannan, and Baker (1996); Raudenbush (1997); and Raudenbush and Liu (2000). Both Orr (1999) and Raudenbush and Liu (2000) address tradeoffs between power and the cost of adding more participants within and between treatment sites.

Often resources will prevent the researcher from including the number of higher order units that the power analyses suggest is required to conduct a sensitive statistical analysis. In such situations, it helps to treat the study as if it were a quasi-experiment, adding features such as switching replications or double pretests to facilitate causal inferences. Shadish, Cook, and Houts (1986) discuss this strategy and provide illustrations. For example, in the Cali study, McKay et al. (1978) staggered the introduction of treatment across the five treatment groups, so that some received treatment for the full length of the study but others received treatment progressively later. All had a common final posttest time. Their demonstration that effects started concurrent with implementation of treatment in each group helped to bolster the study's interpretability despite its use of only four higher order units per condition. Finally, measurement of the characteristics of higher order units helps diagnose the extent to which those characteristics are confounded with treatment.

Researchers sometimes create an unnecessary unit of analysis problem when, in order to save the extra costs and logistical complexity of treating participants individually, they administer to a group a treatment that could have been administered to individuals. By doing this, the researcher may thus create dependencies among participants within groups. For example, suppose a treatment for insomnia is administered to 50 participants in 10 groups of 5 people each; and suppose further that the treatment could have been administered individually in the sense that it does not involve transindividual theoretical components such as mutual interpersonal support. Nonetheless, group members are now exposed to many common influences. For example, some of them might become romantically involved, with possible consequences for their sleep patterns! These group influences may vary from group to group and so affect outcome differentially. So researchers should administer the treatment to individual units if the research question makes this possible; if not, then group membership should be taken into account in the analysis.

The Limited Reach of Random Assignment

Though random assignment is usually better than other design features for inferring that an observed difference between treatment and control groups is due to some cause, its applicability is often limited. Random assignment is useful only if a researcher has already decided that a local molar causal inference is of most interest. Such inferences are a common goal in social research, but they are not the only goal. Yet random assignment is conceptually irrelevant to all other research goals. Further, random assignment is just one part of experimental design, and experimental design is only part of an overall research design. Experimental design involves the scheduling of observations, the choice of treatments and comparisons, the selection of observations and measures, the determination of who should be the respondents, and the manner of assigning units to treatments. Ran-

dom assignment deals with only the last of these issues, so to assign at random does not guarantee a useful experimental or research design.

Thus, if a randomized experiment is conducted with units that do not correspond to the population of theoretical or policy interest, the usefulness of the research is weakened even if the quality of the causal inference is high. Rossi and Lyall (1976) criticized the New Jersey Negative Income Tax Experiment because the respondents were working poor, but most guaranteed incomes in a national scheme would go to the jobless poor. Similarly, Cook et al. (1975) criticized Ball and Bogatz (1970) for manipulating levels of social encouragement to view "Sesame Street," thus confounding viewing with encouragement. Larson (1976) criticized the Kansas City Patrol Experiment because the amount of police patrolling that was achieved in the high-patrolling condition was not even as high as the average in New York City and because the contrast between high- and low-patrol areas in Kansas City was reduced due to police squad cars crossing atypically often over the low-patrol areas with their lights flashing and sirens screaming. These are all useful criticisms of details from social experiments, though none is a criticism of random assignment itself. Such criticisms have implications for the desirability of random assignment only to the extent that implementing such assignment caused the problems to emerge. This is rarely the case.

SOME DESIGNS USED WITH RANDOM ASSIGNMENT

This section reviews many variants of randomized experimental designs (see Table 8.1; for other variations, see Fleiss, 1986; Keppel, 1991; Kirk, 1982; Winer, Brown, & Michels, 1991). The designs we present are the most commonly used in field research, providing the basic building blocks from which more complex designs can be constructed. This section uses the same design notation as in earlier chapters, except that the letter *R* indicates that the group on that line was formed by random assignment. We place *R* at the start of each line, although random assignment could occur either before or after a pretest, and the placement of *R* would vary accordingly.

The Basic Design

The basic randomized experiment requires at least two conditions, random assignment of units to conditions, and posttest assessment of units. Structurally, it can be represented as:

R	*X*	O
R		O

TABLE 8.1 Schematic Diagrams of Randomized Designs

The Basic Randomized Design Comparing Treatment to Control

R	X	O
R		O

The Basic Randomized Design Comparing Two Treatments

R	X_A	O
R	X_B	O

The Basic Randomized Design Comparing Two Treatments and a Control

R	X_A	O
R	X_B	O
R		O

The Pretest-Posttest Control Group Design

R	O	X	O
R	O		O

The Alternative-Treatments Design with Pretest

R	O	X_A	O
R	O	X_B	O

Multiple Treatments and Controls with Pretest

R	O	X_A	O
R	O	X_B	O
R	O	O	

Factorial Design

R	X_{A1B1}	O
R	X_{A1B2}	O
R	X_{A2B1}	O
R	X_{A2B2}	O

Longitudinal Design

R	$O \ldots O$	X	O	$O \ldots O$
R	$O \ldots O$		O	$O \ldots O$

A Crossover Design

R	O	X_A	O	X_B	O
R	O	X_B	O	X_A	O

Note: For simplicity, we place the *R* (to indicate random assignment) at the front of the schematic diagram; however, assignment sometimes occurs before and sometimes after the pretest, so that the placement of *R* could be varied accordingly.

A good example of the use of this design with a single treatment and a control group is the test of the Salk polio vaccine in 1954. More than 400,000 children were randomly assigned to receive either the vaccine or a placebo (Meier, 1972).

A key issue is the nature of the control condition. Selection of a particular kind of control group depends on what one wants to control. For example, a no-treatment control condition tests the effects of a molar treatment package, including all its active and passive, important and trivial components. However, when interest is in the effects of a part of that package, the control should include everything but that part. In drug studies, for example, the researcher often wants to separate out the effects of the pharmaceutically active ingredients in the drugs from the effects of the rest of the package—things such as swallowing a pill or having contact with medical personnel. A placebo control does this, with medical personnel providing patients with, say, an inert pill in a manner that includes all the extraneous conditions except the active ingredients (Beecher, 1955).

Many types of control groups exist, for example, no-treatment controls, dose-response controls, wait-list controls, expectancy controls, or attention-only controls (Borkovec & Nau, 1972; Garber & Hollon, 1991; International Conference on Harmonization, 1999; Jacobson & Baucom, 1977; Kazdin & Wilcoxon, 1976; O'Leary & Borkovec, 1978; Orne, 1962; Seligman, 1969; Shapiro & Shapiro, 1997). The variations are limited only by the researcher's imagination. But in all cases the question is always, "Control for what?" For instance, Rossi and Lyall (1976, 1978) criticized the New Jersey Negative Income Tax in part on this basis—that the control group differed not only in failure to receive the treatment of interest but also in receiving far fewer and less intrusive administrative experiences than the treatment group.

Two Variants on the Basic Design

One variation compares two treatments by substituting X_A and X_B for the X and blank space in the previous diagram:

$$\begin{array}{lll} R & X_A & O \\ R & X_B & O \end{array}$$

If X_A is an innovative treatment, for example, X_B is often a "gold-standard" treatment of known efficacy. The causal question is then, "What is the effect of the innovation compared with what would have happened if units had received the standard treatment?" This design works well if the standard treatment has a known track record against no-treatment controls. But if not, and if those receiving X_A are not different from those receiving X_B at posttest, the researcher cannot know if both treatments were equally effective or equally ineffective. In that case, a control group helps:

$$\begin{array}{lll} R & X_A & O \\ R & X_B & O \\ R & & O \end{array}$$

This design was used in Boston to study the effects of an experimental housing project designed to improve the kinds of neighborhoods in which poor families lived (Katz, Kling, & Liebman, 1997; Orr, 1999). Poverty families in Treatment A received housing vouchers good for use only in low-poverty areas so that if they moved, they would move to better neighborhoods; those in Treatment B received vouchers for use anywhere, including in high-poverty areas; and those in the control group did not receive any vouchers at all.

Risks to This Design Due to Lack of Pretest

Omitting a pretest is a virtue whenever pretesting is expected to have an unwanted sensitization effect; and it is a necessity when a pretest cannot be gathered (as in some studies of cognitive development in infants), is seriously impractical (as with expensive and time-consuming interviews of patients by physicians), or is known to be a constant (as in studies of mortality in which all patients are alive at the start). Otherwise, the absence of a pretest is usually risky if there is any likelihood of attrition from the study; in fact, some observers cite the need for a pretest as one of the most important lessons to emerge from the last 20 years of social experiments (Haveman, 1987). Attrition occurs often in field experiments, leaving the researcher with the need to examine whether (1) those who dropped out of the study were different from those who remained and, especially, (2) if those who dropped out of one condition were different from those who dropped out of the other condition(s). Pretreatment information, preferably on the same dependent variable used at posttest, helps enormously in answering such questions.

Of course, attrition is not inevitable in field experiments. In medical trials of surgical procedures that have immediate outcomes, the treatments happen too quickly to allow much attrition; and patient follow-up care is often thorough enough and medical records good enough that posttests and follow-up observations on patients are available. An example is Taylor et al.'s (1978) study of short-term mortality rates among 50 heart attack patients randomly assigned to receive either manual or mechanical chest compression during cardiopulmonary resuscitation. The intervention was started and finished within the space of an hour; the heart attack patients could not very well get up and leave the hospital; and the dependent variable was quickly and easily gathered. A second situation conducive to minimal attrition is one in which the outcome is a matter of mandatory public record. For instance, in both the LIFE (Living Insurance for Ex-Offenders) experiment and the TARP (Transitional Aid for Released Prisoners) experiment (Rossi, Berk, & Lenihan, 1980), the main dependent variable was arrests, about which records were available for all participants from public sources. In general, however, attrition from conditions will occur in most field experiments, and pretests are vital to the methods we outline in Chapter 10 for dealing with attrition.

The Pretest-Posttest Control Group Design

Consequently, adding pretests to the basic randomized design is highly recommended:

$$\begin{array}{cccc} R & O & X & O \\ R & O & & O \end{array}$$

Or, if random assignment occurred after pretest,

$$\begin{array}{cccc} O & R & X & O \\ O & R & & O \end{array}$$

This is probably the most commonly used randomized field experiment. Its special advantage is its increased ability to cope with attrition as a threat to internal validity, in ways we outline in Chapter 10. A secondary advantage, however, is that it allows certain statistical analyses that increase power to reject the null hypothesis (Maxwell & Delaney, 1990). S. E. Maxwell (1994) says that allocating 75% of assessment to posttest and 25% to pretest is often a good choice to maximize power with this design. Maxwell, Cole, Arvey, and Salas (1991) discuss tradeoffs between ANCOVA using a pretest as a covariate and repeated-measures ANOVA with a longer posttest as a method for increasing power.

Although the researcher should try to make the pretest be identical to the outcome measures at posttest, this need not be the case. In research on child development, for example, tests for 8-year-old children must often be substantially different in content than those for 3-year-old children. If the pretest and posttest assess the same unidimensional construct, logistic test theory can sometimes be used to calibrate tests if they contain some common content (Lord, 1980), as McKay et al. (1978) did in the Cali study of changes in cognitive ability in 300 children between the ages of 30 and 84 months.

Alternative-Treatments Design with Pretest

The addition of pretests is also recommended when different substantive treatments are compared:

$$\begin{array}{cccc} R & O & X_A & O \\ R & O & X_B & O \end{array}$$

If posttests reveal no differences between groups, the researcher can examine pretest and posttest scores to learn whether both groups improved or if neither did.[5] This design is particularly useful when ethical concerns mitigate against comparing treatment

5. Here and elsewhere, we do not mean to imply that change scores would be desirable as measures of that improvement. ANCOVA will usually be much more powerful, and concerns about linearity and homogeneity of regression are at least as important for change scores as for ANCOVA.

with a control condition, for example, in medical research in which all patients mus be treated. It is also useful when some treatment is the acknowledged gold standar against which all other treatments must measure up. Comparisons with this standar treatment have particularly practical implications for later decision-making.

Multiple Treatments and Controls with Pretest

The randomized experiment with pretests can involve a control group and multi ple treatment groups:

R	O	X_A	O
R	O	X_B	O
R	O		O

H. S. Bloom's (1990) study of reemployment services for displaced worker used this design. More than 2,000 eligible unemployed workers were assigned ran domly to job-search assistance, job-search assistance plus occupational training, o no treatment. Note that the first treatment included only one part of the treatmen in the second condition, giving some insight into which parts contributed most t outcome. This is sometimes called a **dismantling study**, though the dismantling wa only partial because the study lacked an occupational-training-only condition Clearly, resources and often logistics prevent the researcher from examining too many parts, for each part requires a large number of participants in order to test i well. And not all parts will be worth examining, particularly if some of the indi vidual parts are unlikely to be implemented in policy or practice.

This design can be extended to include more than two alternative treatments o more than one control condition. An example is the National Institute of Menta Health Treatment of Depression Collaborative Research Program (NIMH-TDCRP; Elkin, Parloff, Hadley, & Autry, 1985; Elkin et al., 1989; Imber et al., 1990). In this study, 250 depressed patients were randomly assigned to receive cognitive behavio therapy, interpersonal psychotherapy, antidepressant chemotherapy (imipramine) plus clinical management, or a placebo pill plus clinical management.

This design is also used to vary the independent variable in a series of increas ing levels (sometimes called *parametric* or *dose-response* studies). For example, the Housing Allowance Demand Experiment randomly assigned families to receive housing subsidies equal to 0%, 20%, 30%, 40%, 50%, or 60% of their rent (Friedman & Weinberg, 1983). The Health Insurance Experiment randomly as signed families to insurance plans that required them to pay 0%, 25%, 50%, or 95% of the first $1,000 of covered services (Newhouse, 1993). The more levels of treatment are administered, the finer the assessment can be of the functional form of dosage effects. A wide range of treatment levels also allows the study to detect effects that might otherwise be missed if only two levels of a treatment that are not powerful enough to have an effect are varied. The Cali, Colombia, study (McKay et al., 1978), for example, administered a combined educational, nutritional, and

medical treatment in four increasing dosage levels—990 hours, 2,070 hours, 3,130 hours, and 4,170 hours. At the smallest dosage—which itself took nearly a full year to implement and which might well be the maximum dosage many authors might consider—the effects were nearly undetectable, but McKay et al. (1978) still found effects because they included this wide array of higher dosages.

Factorial Designs

These designs use two or more independent variables (called *factors*), each with at least two levels (Figure 8.1). For example, one might want to compare 1 hour of tutoring (Factor A, Level 1) with 4 hours of tutoring (Factor A, Level 2) per week and also compare tutoring done by a peer (Factor B, Level 1) with that done by an adult (Factor B, Level 2). If the treatments are factorially combined, four groups or cells are created: 1 hour of tutoring from a peer (Cell A1B1), 1 hour from an adult (A1B2), 4 hours from a peer (A2B1), or 4 hours from an adult (A2B2). This is often described as a 2 × 2 ("two by two") factorial design written in the notation used in this book as:

$$\begin{array}{lll} R & X_{A1B1} & O \\ R & X_{A1B2} & O \\ R & X_{A2B1} & O \\ R & X_{A2B2} & O \end{array}$$

This logic extends to designs with more than two factors. If we add a third factor in which the tutor is or is not trained in effective tutoring methods (Factor C, Levels 1 and 2), we have a 2 × 2 × 2 design with 8 possible cells. The levels of the factors can include control conditions, for example, by adding a no-tutoring condition to Factor A. This increases the number of levels of A so that we have a 3 × 2 × 2 design with 12 cells. This notation generalizes to more factors and more levels in similar fashion.

Factorial designs have three major advantages:

- They often require fewer units.
- They allow testing combinations of treatments more easily.
- They allow testing interactions.

First, they often allow smaller sample sizes than would otherwise be needed.[6] An experiment to test for differences between peer versus adult tutoring might require

6. Two exceptions to this rule are (1) detecting interactions of special substantive interest may require larger sample sizes because power to detect interactions is usually lower than power to detect main effects; and (2) if the outcome is a low base rate event (e.g., death from pneumonia during the course of a brief clinical trial) and if the treatments in both factors reduce death, their combined effect may reduce the number of outcome events to the point that more participants are needed in the factorial design than if just one treatment were tested.

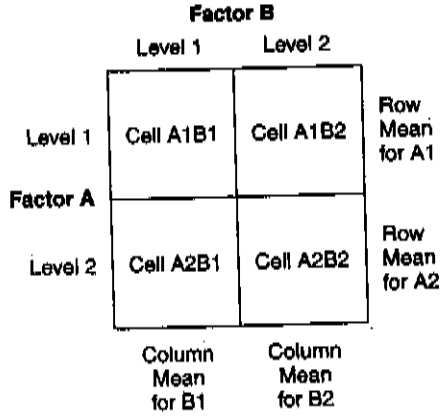

FIGURE 8.1 Factorial Design Terms and Notation

50 participants per condition, as might a second experiment to test for differences between 1 versus 4 hours of tutoring—a total of 200 participants. In a factorial design, fewer than 200 participants may be needed (the exact number would have to be determined by a power analysis) because each participant does double duty, being exposed to both treatments simultaneously.

Second, factorial designs allow the investigator to test whether a combination of treatments is more effective than one treatment. Suppose that an investigator runs both an experiment in which participants are assigned to either aspirin or placebo to see if aspirin reduces migraine headaches and a second experiment to test biofeedback versus placebo for the same outcome. These two experiments provide no information about the effects of aspirin and biofeedback applied jointly. The factorial design provides information about the effects of aspirin only, biofeedback only, aspirin plus biofeedback, or no treatment.

Third, factorial experiments test interactions among factors (Abelson, 1996; D. Meyer, 1991; Petty, Fabrigar, Wegener, & Priester, 1996; Rosnow & Rosenthal, 1989, 1996). Treatments produce *main effects;* for example, the main effect of aspirin relative to a placebo pill is to reduce headaches. Main effects are *average* effects that may be misleading if, for example, some kinds of headaches respond well to aspirin but others do not. Interactions occur when treatment effects are not constant but rather vary over levels of other factors, for example, if aspirin reduces tension headaches a lot but migraine headaches very little. Here the treatment (aspirin) interacts with a moderator variable (type of headache), the word *moderator* describing a second factor that interacts with (moderates the effect of) a treatment. The same general logic extends to designs with three or more factors, though higher order interactions are more difficult to interpret.

Interactions are often more difficult to detect than are main effects (Aiken & West, 1991; Chaplin, 1991, 1997; Cronbach & Snow, 1977; Fleiss, 1986), so

large sample sizes with appropriate power analyses are essential whenever interactions are an important focus.[7] Indeed, some authors have argued that predicted interactions are sufficiently important in advancing scientific theory to warrant testing them at a larger than usual Type I error rate (Meehl, 1978; Platt, 1964; Smith & Sechrest, 1991; Snow, 1991). If testing a predicted interaction is at issue, deliberate oversampling of observations that are extreme on the interacting variables provides a more powerful (and still unbiased) test of the interaction, although it gives a poorer estimate of the total variance accounted for by the predictors. The test for the interaction could be done using an unweighted sample, and the test for total variance could be done using a sample that is weighted to reflect the population of interest (McClelland & Judd, 1993). This is a special case of optimal design theory (e.g., A. Atkinson, 1985) that can help select treatment levels and combinations that maximize the power of the design to detect parameters that may be of particular policy or theoretical interest.

When using a factorial design, the researcher need not actually assign units to all possible combinations of factors, though empty cells can reduce power. It might waste resources to test treatment combinations that are of no theoretical interest or unlikely to be implemented in policy. The New Jersey Negative Income Tax (NJNIT) experiment, for example, studied proposals for dealing with poverty and welfare reform (Kershaw & Fair, 1976)—specifically, the joint effects of two independent variables: the guarantee level and the tax rate. Guarantee level was an amount of money paid to poor families or individuals if they had no other income; it was defined as 50%, 75%, 100%, or 125% of the poverty level. The tax rate was the rate at which that guaranteed income was reduced as a family's other income rises: 30%, 50%, or 70%. So the design was a 4×3 factorial experiment that could assign participants to 12 different cells. However, the investigators assigned the 725 participants only to the eight cells that were not too costly and that were considered politically feasible for eventual policy implementation. The empty cells can complicate data analysis, but the flexibility of this option often outweighs the complications. This design is an example of a fractional factorial design that allows estimates of some higher order interaction terms even when the full factorial design is not implemented (Anderson & McLean, 1984; Box, Hunter, & Hunter, 1978; West, Aiken, & Todd, 1993).

Nested and Crossed Designs

In a crossed design, each level of each factor is exposed to (crossed with) all levels of all other factors. For example, in an educational experiment, if some students in each classroom are exposed to treatment and some to control, then the

7. Interactions are ordinal (when you graph the cell means, the resulting lines do not cross) or disordinal (they do cross) (Maxwell & Delaney, 1990). Tests of ordinal interactions frequently have lower power than main effects, but tests of disordinal interactions are often more powerful than the test of either main effect. Rosnow and Rosenthal (1989) explain which lines should and should not cross when interactions are present.

treatment factor is crossed with classroom. In a **nested design**, some levels of one factor are not exposed to all levels of the other factors. For example, when some classrooms receive the treatment but not the control condition, classrooms are nested within treatment conditions. Crossed designs yield unconfounded statistical tests of all main effects and interactions, but nested designs may not. The distinction between nested and crossed designs is particularly relevant in the presence of higher order units (e.g., schools, hospitals, work sites). Often the researcher will nest treatments within these units to minimize the chances of diffusion and communication of treatment within higher order units. The dilemma is that the crossed design yields separate statistical estimates of the effects of higher order units, treatment conditions, and their interaction; but crossing increases problems such as diffusion of treatment. Each researcher will have to review the specifics of this tradeoff as it applies to the experiment at hand before deciding whether nesting or crossing is to be preferred.

A Disadvantage of Factorial Designs

Factorial designs are common in laboratory research and in highly controlled settings, such as those of some medical research. They are more difficult to implement in many field settings. They require close control over the combination of treatments given to each unit, but such control is difficult as more factors or more levels are included—especially if each cell has different eligibility criteria, as in pharmaceutical studies in which rules for who can receive which drug combinations can be complex. In addition, much field research is conducted to assess the policy implications of a proposed innovation. Yet the ability of policymakers to legislate or regulate interactions is low, given traditions of local control and professional discretion in the delivery of services and given difficulties in ensuring that social interventions are implemented as intended (Pressman & Wildavsky, 1984; Rossi & Wright, 1984). Policymakers are often more interested in generalized inferences about which treatments work than in the highly specific and localized inferences about the effects of particular combinations of particular levels of particular factors in a particular setting that factorial designs sometimes provide.

Longitudinal Designs

Longitudinal designs add multiple observations taken before, during, or after treatment, the number and timing of which are determined by the hypotheses under study:

$$\begin{array}{llll} R & O \ldots O & X & O \quad O \ldots O \\ R & O \ldots O & & O \quad O \ldots O \end{array}$$

These designs closely resemble the time-series studies in Chapter 6, but they have far fewer pre- and posttest observations. Longitudinal designs allow examination

of how effects change over time, allow use of growth curve models of individual differences in response to treatment, and are frequently more powerful than designs with fewer observations over time, especially if five or more waves of measurement are used (Maxwell, 1998). So especially when sample sizes are small, adding pretests and posttests can improve power.

Longitudinal randomized experiments with multiple pretests are rare. Bloom (1990), for example, randomly assigned displaced workers to three treatment or control conditions designed to help them get jobs. He reported quarterly earnings at four pretests and four posttests. The pretests showed that participants experienced an acute drop in earnings during the one or two quarters immediately preceding assignment to conditions, perhaps reflecting a short-term job loss by workers who move rapidly into and out of the labor market. So regression effects might cause some improvement in all groups even if treatments did not work. Indeed, control participants did improve, though not as much as treatment participants.

The use of multiple posttests is more common. For example, the Cambridge-Somerville Youth Study began in 1939 when 650 adolescent boys were randomly assigned from blocked pairs either to a counseling program or to no treatment (W. McCord & McCord, 1959; Powers & Witmer, 1951), with a follow-up taken 37 years later in 1976 (J. McCord, 1978). A current study in a health maintenance organization aims to follow patients for their entire lives (Hillis et al., 1998). Here the multiple posttests explore whether treatment gains are maintained or changed over time. This is especially important if the primary outcome can be measured only many years later—for example, children's eventual educational and occupational achievement after participating in Head Start, mortality rates from AIDS among gay men after exposure to a program teaching safe sex, or lifetime earned income among Job Corps trainees. Sometimes longitudinal studies follow different outcomes simultaneously over time to explore the validity of a hypothesized causal chain of effects—for example, that a treatment to help children of lower socioeconomic status to rise out of poverty will first improve their aspirations, which will affect expectations, which will affect achievements in grammar school, which will help them successfully complete high school and college, which will finally lead to a better paying or higher status job as an adult. Here the timing of the observations follows the hypothesized schedule of events to be observed in order to explore if and at what point the chain breaks down.

Practical problems plague longitudinal designs. First, attrition rises with longer follow-up periods as, for example, participants move to unknown locations or simply tire of the research. Still, we are impressed with what tireless follow-up procedures can achieve with most populations (Ribisl et al., 1996). Second, some long-term outcomes, such as lifetime earned income, are nearly impossible to assess given current technology and limited access to such relevant data sources as the Internal Revenue Service or the Social Security Administration (Boruch & Cecil, 1979). Third, it is not always ethical to withhold treatments from participants for long periods of time, and the use of longitudinal observations on no-treatment or wait-list control-group participants is rare because such

participants often simply obtain treatment elsewhere. An example of all these problems is provided by Snyder and Wills (1989; Snyder, Wills, & Grady-Fletcher, 1991), who randomly assigned 79 distressed couples to receive behavioral marital therapy ($N = 29$), insight-oriented marital therapy ($N = 30$), or a wait-list control group ($N = 20$). At 6-month and 4-year follow-ups they assessed outcomes only on the two treatment group conditions because control participants had already begun dropping out of the study despite the agreed-upon 3-month waiting period between pretest and posttest. Despite participant death, medical problems, and relocation out of state, Snyder and Wills (1989) were able to gather 4-year follow-up data on 55 of the 59 treatment couples—a remarkably high retention rate, although still a loss of participants such as one nearly always experiences in longitudinal research. Finally, a 4-year follow-up is far longer than most psychotherapy outcome studies use, but it is still far short of such long-term outcomes as distress levels over the life of the marriage or lifetime divorce rates. Even exemplary longitudinal studies such as this experience such problems.

Crossover Designs

Imagine an experiment in which some participants are randomly assigned to receive either Treatment A or B, after which they receive a posttest. In a crossover design, after that posttest the participants cross over to receive the treatment they did not previously get, and they take another posttest after that second treatment is over. In our design notation this crossover design is written:

R	O	X_A	O	X_B	O
R	O	X_B	O	X_A	O

Sometimes the interval between treatments is extended so that the effects of the first treatment can dissipate before the second treatment begins.

This design is often used in medical research, as in cases in which several drugs are given to participants in a within-participants design and the crossover is used to counterbalance and assess order effects.[8] It is also used to gather even more causal information from a study that would otherwise stop after the first posttests were administered. In either use, the crossover design is most practical when the treatments promise short-term relief (otherwise carryover effects will occur), when the treatments work quickly (otherwise the experiment will take too long), and when participants are willing and able to continue through both treatments even if the first treatment fixes the problem. If analysis finds an interaction between treatments and

8. The crossover design is a variation of a more general class called Latin squares (Cochran & Cox, 1957; Fisher & Yates, 1953; Fleiss, 1986; R. Kirk, 1982; Pocock, 1983; Rosenthal & Rosnow, 1991; Winer et al., 1991). Latin squares are widely used to counterbalance treatments in within-participants factors and to estimate effects in very large factorial designs in which all possible combinations of conditions cannot be administered.

TABLE 8.2 Ten Situations Conducive to Randomized Experiments

1. When demand outstrips supply
2. When an innovation cannot be delivered to all units at once
3. When experimental units can be temporally isolated
4. When experimental units are spatially separated or interunit communication is low
5. When change is mandated and solutions are acknowledged to be unknown
6. When a tie can be broken or ambiguity about need can be resolved
7. When some persons express no preference among alternatives
8. When you can create your own organization
9. When you have control over experimental units
10. When lotteries are expected

order, then the effect of the second round of treatments cannot be interpreted without taking order effects into account, although the first round of treatment is still just as interpretable as would have been the case without the crossover.

CONDITIONS MOST CONDUCIVE TO RANDOM ASSIGNMENT

This section (and Table 8.2) explicates the situations that increase the probability of successfully doing a randomized field experiment.

When Demand Outstrips Supply

When demand for service outstrips supply, randomization can be a credible rationale for distributing service fairly. For example, Dunford (1990) describes an experiment on the effects of a summer youth employment program. Initially, program personnel objected to randomly assigning some youths to jobs and others not. However, they also recognized that far fewer jobs were available than there were applicants, and they eventually agreed that random allocation of those jobs was fair. They later reported that the obviously unbiased nature of randomization helped them to show a vocal group of critics that entry into the program discriminated neither for nor against minority youth. Similarly, the Omnibus Budget Reconciliation Act of 1981 allowed states to experiment with novel approaches to welfare reform. Many states wanted to do so, but few states could afford to implement programs that could be given to all welfare recipients; random assignment was again accepted as a fair mechanism for distributing services in one experiment

(Gueron, 1985). Finally, the Milwaukee Parental Choice Program tested the use of school vouchers by random selection of participants when there were more applicants to a particular school and grade than could be accommodated (Rouse, 1998).

When demand exceeds supply, applicants originally assigned to the comparison condition sometimes reapply for the treatment. Experimenters need to be clear about whether they will have this right, and if they do, whether reapplicants will have priority over new applicants. Sometimes the right to reapply cannot be denied on ethical or regulatory grounds, as in the case of a distressed psychotherapy client assigned to a wait-list who becomes severely symptomatic or that of welfare recipients who have a regulatory right to reapply to a job-training program. It is crucial to negotiate support from everyone in the experiment about dealing with reapplicants, for dissenters can thwart those arrangements (Conrad, 1994). For example, the Rockefeller Foundation's Minority Female Single Parent (MFSP) program could not afford to provide services to all eligible candidates right away, so randomization was proposed as an ethically appropriate way to distribute services among the many eligible women (Boruch, 1997). However, some local program managers disagreed and spread their resources more thinly over a large number of women rather than limit the number of women served. Ultimately, if a large proportion of rejected applicants is likely to reapply and be accepted into treatment, the feasibility of a randomized experiment is questionable. If the proportion of successful reapplicants is likely to be small, methods that we discuss in Chapter 10 for dealing with treatment implementation problems may be useful.

When an Innovation Cannot Be Delivered to All Units at Once

Often it is physically or financially impossible to introduce an innovation simultaneously to all units. Such situations arise in education as curricula are slowly changed, as new teaching devices filter down through the schools in a system, or as computers are introduced or new training schemes are implemented. In these situations, the experiment can deliberately introduce the innovation in stages, with some units receiving it before others on a random basis. This provides an experimental and control comparison until the point at which the controls get their turn for treatment. It is even better if it can be done using the switching-replications design feature described for previous quasi-experimental designs, but with replications now randomly assigned.

When Experimental Units Can Be Temporally Isolated: The Equivalent-Time-Samples Design

Although we typically think of randomly assigning people, schools, communities, or cities to conditions, we can also randomly assign times to conditions (Hahn,

1984). Campbell and Stanley (1963) called this an "Equivalent Time Samples Design" to highlight that randomization equates the time periods in which the treatment was present to those in which it was absent. Edgington (1987) provides several examples of single-participant designs in which treatments were presented and removed randomly over time—a comparison of three drugs for narcolepsy, of a medication with a placebo for an intestinal disorder, and of the effects of artificial food colorings with placebo on the behavior of hyperactive children. The effect must be of short duration so that it can decrease in magnitude when treatment is withdrawn; and the effect must continue to respond to repeated exposure to treatment so that it can increase when treatment is readministered.

But the principle applies to more than just single-participant designs. It can be used when there are naturally occurring rotations of groups and each group is isolated from the others in time. Thus, when 24 groups of persons came for sequential 2-week stays at a pastoral counseling center, Mase (1971) randomly assigned those groups to one of two kinds of sensitivity training, twelve groups receiving each kind. In this example, the creation of simultaneous treatment and control conditions might have led to diffusion of treatment or other reactive threats to validity, but the equivalent-time-samples design avoided such problems. Note, however, that participants are now nested within time samples in the same way they could be nested within some aggregate such as a school or a neighborhood; the analysis should take this into account.

When Experimental Units Are Spatially Separated or Interunit Communication Is Low

When units are geographically separated and have minimal contact with one another, or when they can be made this way, those units can be randomly assigned. This often occurs in organizations that have many branches, for example, supermarkets, units in the armed forces, university alumni, schools within school districts, wards within hospitals, residential units of religious orders, branches of health clubs in large cities, and dealerships that sell automobiles, appliances, and the like. However, spatial isolation does not guarantee minimal contact, so care should be taken to check that this is indeed the case.

For example, an experiment in Peru studied the effects of providing gynecological and family planning services to clients of 42 geographically separated community clinics (Population Council, 1986). Clinics were assigned randomly to receive one, two, or four physician visits per month. The geographical separation of clinics meant that women tended to visit the same clinic over time, so little diffusion of treatment was likely. If some diffusion was possible (e.g., if women regularly visited two clinics very close to each other), the researchers could have blocked clinics by geographic area and assigned areas rather than individual clinics. Similarly, Perng (1985) randomly assigned people to six different methods that the Internal Revenue Service was considering for collecting delinquent income tax

returns. Most people were separated geographically. But even if they had been in close physical proximity to each other, by law the very fact that their tax return was part of a study was confidential, and people are generally reluctant to discuss their income tax returns; so it was unlikely that communication between people in different conditions would occur.

These experiments had an additional strength; both took advantage of the natural appearance of the interventions to randomize treatments unobtrusively. After all, patients expect that physicians will visit clinics and are not likely to notice minor variations in the number of times those visits occur. Those receiving delinquent tax letters from the IRS are rarely familiar enough with specific IRS procedures to know that any variation on normal routine was occurring. Unobtrusiveness is a worthy goal to strive for, except when treatment is deliberately designed to stand out from what respondents expect.

When Change Is Mandated and Solutions Are Acknowledged to Be Unknown

Sometimes, all concerned parties agree that an undesirable situation needs changing, but it is not clear which changes we should make despite passionate advocacy of certain alternatives by interested parties. If administrative, political, and economic conditions allow, trying out several alternative changes in a formal experiment is more likely to win acceptance. An example is the Minneapolis Spouse Abuse Experiment (Berk, Smyth, & Sherman, 1988). Spouse abuse is a serious felony that can lead to the murder of the spouse, and so police officers who are called to such a crime must take some action. But concerned parties disagreed about whether that action should be to do on-the-spot counseling between the two spouses, to require the offender to leave the premises for 8 hours, or to arrest the offender. An administrator who had an attitude favoring experimentation in finding a solution allowed the implementation of a randomized experiment to test which of these three options worked best. Similarly, a randomized experiment to treat severely mentally ill patients with either standard care or a radically different form of community care could be done in part because all parties acknowledged that they were unsure which treatment worked best for these patients (Test & Burke, 1985).

Though such planned variation studies promise important results, each variation may not define its goals *exclusively* in terms of the same target problem. In the Minneapolis Spouse Abuse Experiment, this potential disagreement was not a problem because most parties agreed that the end point of interest was a decrease in postintervention repeat violence. However, disagreement may be more likely if participants are assigned to projects with different management, staff, and funders than to projects in which all variations are implemented by the same people. Nor will the directors of the various projects always agree which measures should be used to measure those things they are trying in common to change.

When a Tie Can Be Broken or Ambiguity About Need Can Be Resolved

Assignment of people to conditions on the basis of need or merit is often a more compelling rule to program managers, staff, and recipients than is randomization. Such considerations are one justification for the regression discontinuity design. However, the need or merit of some people is often ambiguous. In those cases, the ambiguity can sometimes be resolved by randomly assigning people of ambiguous need to conditions, perhaps in combination with the regression discontinuity design. Similarly, Lipsey, Cordray, and Berger (1981) used random assignment to resolve ambiguity in their evaluation of a juvenile delinquency diversion program. In a quasi-experimental design, police officers used their best judgment as to whether an arrested juvenile needed to be counseled and released, referred to probation, or diverted to a more intensive social service project that provided counseling, remedial education, recreation, and substance abuse services. However, when the officer was unsure which assignment was most needed and also judged that either counseling and release or diversion would be appropriate, the officer randomized juveniles to one of these two conditions.

In such tie-breaking experiments, generalization is restricted to persons scoring in the area of ambiguous need, the group about which we know least as far as effective treatment. However, if an organization specializes in treating the best, the worst, or the full range of participants, its officials may well object that evaluating their performance with "ambiguous" participants is insensitive to what they really do. Fortunately, it may be possible to link a tie-breaking experiment with some form of interpretable quasi-experiment, as Lipsey et al. (1981) did, to satisfy these objections.

When Some Persons Express No Preference Among Alternatives

Even if ethics or public relations require that people be allowed to choose which option they will receive, persons who express no preference from among the options can be assigned by chance. For example, Valins and Baum (1973) wanted to study some effects of physical environment on university freshmen who entered one of two kinds of living quarters that differed in the number of persons a resident was likely to meet each day. The authors restricted the study to the 30% of freshmen who expressed no preference for either kind of living quarters. College authorities assigned this 30% to living units on a haphazard basis; but it would presumably have been easy to do the assignment randomly. Of course, limiting the experiment to persons who have no preference does make generalization beyond such persons more problematic. If the full range of decisive and no-preference respondents is of interest, the randomized experiment with the no-preference respondents

could be conducted along with the best possible quasi-experiment with the decisive respondents. Then the results of the studies can be compared, with the weakness of one study being the strength of the other. Where the results coincide, a global overall inference is easier.

When You Can Create Your Own Organization

Random assignment is an accepted part of the organizational culture of laboratory experimentation, but most field experiments are conducted in organizational cultures in which randomization is mostly foreign. Yet sometimes researchers can create their own organizations in which they can make the practice of randomization a more usual norm. For example, university psychology departments often set up a psychological services center to facilitate the training of graduate students in clinical psychology and to allow department faculty members to exert more experimental control than would typically be possible in most clinics (Beutler & Crago, 1991). In such centers, researchers can better control not just randomization but also such features as treatment standardization, measurement, and case selection. Freestanding research institutes and centers focused on particular problems frequently allow similar levels of broad control. The California Smokers' Helpline, for example, provides free smoking cessation help to smokers in that state who call the helpline (Zhu, 1999). Randomizing callers to treatment and control was not feasible. All callers received a treatment mailing with instructions to call back when they were ready to start treatment. Those who did not call back were then randomized into two groups: no further action or proactive callback from the treatment staff to begin treatment. In principle, this procedure could be used to randomly subdivide the nonresponders in any quasi-experimental treatment group into treatment and control—for example, those who request psychotherapy but fail to show for appointments, those who are given prescriptions but fail to fill them, those who are accepted to a job training program but fail to attend, and so forth. Finally, researchers can sometimes set up organizations just to control randomization, as is often done in multisite medical trials in which a central clearinghouse controlled by the researcher is created to do randomization. The National Institute of Mental Health (NIMH) Collaborative Depression Project (Collins & Elkin, 1985) used this clearinghouse method to control randomization.

When You Have Control over Experimental Units

Being able to establish one's own organization or randomization clearinghouse is rare. Most field researchers are guests in someone else's organization, and they derive many of their possibilities for control from their powerful hosts. An example comes from an evaluation of solutions to the "peak load" problem by utility com-

panies (Aigner & Hausman, 1980). Electricity usage varies by time of day, and the utility company must have enough capacity to meet peak demand even if that capacity is largely unused at other times. Building that capacity is expensive. So utility companies wanted to know whether charging more for electricity during peak demand periods would reduce demand and so reduce the need to build more capacity. Experimenters were able to randomly assign households to higher peak demand rates versus standard rates because their hosts completely controlled electricity supply to the affected households and were interested in getting an answer to this question with experimental methods.

Randomization is also more likely whenever major funders insist on it. For example, both the National Institute on Drug Abuse and the National Institute on Alcohol and Alcohol Abuse have offered funding for innovative service provision contingent upon evaluation of those services by rigorous experimental designs, both paid for by the grant (Coyle, Boruch & Turner, 1991). The NIMH Collaborative Depression project used the same approach (Boruch & Wothke, 1985). However, especially when funder and fundee have a long-term relationship, the use of the purse strings for control can lead to tension. Lam, Hartwell, and Jekel (1994), for example, noted the "contentious codependence" (p. 56) that developed between Yale University and the city of New Haven, in which Yale is located, due to the fact that Yale researchers frequently offer social services to the city that it might not otherwise be able to afford but with a research string attached. There is a thin line between contentious codependence and oblique coercion, and it is even self-defeating to conduct a randomized experiment in a way that directly or indirectly demeans hosts or respondents. After all, the motivation for hosts and participants to volunteer tomorrow may well be related to how we treat them in experiments today.

When Lotteries Are Expected

Lotteries are sometimes used as a socially accepted means of distributing resources. Examples include a lottery used to assign female students to dormitories at Stanford (Siegel & Siegel, 1957), a lottery to choose among applicants to a newly developed "magnet" school (Zigulich, 1977), and the 1970 draft lottery in the United States (Notz, Staw, & Cook, 1971). In the latter case, Hearst, Newman, and Hulley (1986) asked whether being randomly assigned an eligible draft number elevated mortality and found that it did do so. Angrist et al. (1996a) confirmed this finding, with the average causal effect on mortality of being randomly assigned an eligible draft number equal to less than one tenth of one percent. In these cases, the motivation for randomization was not to do research but rather to capitalize on the perception that randomization is an unbiased way of distributing a resource. These social uses of randomization create a natural randomized experiment that the investigator can exploit. Unfortunately, formal social lotteries do not occur frequently, so they cannot be relied upon as a means of creating probabilistically equivalent groups very often.

WHEN RANDOM ASSIGNMENT IS NOT FEASIBLE OR DESIRABLE

Even when interest exists in whether a treatment is effective, some circumstances mitigate against using a randomized experiment to answer the question. First, randomized experiments may not be desirable when quick answers are needed. Typically, several years pass between the conception of a major field experiment and the availability of results—particularly if the treatment requires time (as with long-term psychotherapy) and if medium- to long-term outcomes are of interest (as with lifetime earnings). In the New Jersey Negative Income Tax Experiment, for example, "the four years of the operating phase were sandwiched between 44 months of planning and design and 16 months of data analysis" (Haveman, 1987, p. 180)—8 years total. So, if information is needed rapidly, alternatives to randomized experiments may be better. For example, the Program Evaluation and Methodology Division (PEMD) of the U.S. General Accounting Office (GAO) frequently fielded questions from legislators who wanted answers quickly about pending decisions. Some of those questions involved the effects of programs or policies. A delay of a few years might delay the decision too long—indeed, the question may no longer be of policy interest, and the legislator who asked the question may no longer be serving. Consequently, PEMD rarely used randomized experiments, relying instead on combinations of quasi-experiments, surveys, and reviews of existing literature about the effects of related policies (Chan & Tumin, 1997; Datta, 1997; Droitcour, 1997). Such procedures may be weaker for inferring cause than a new randomized experiment, because even when the literature contains randomized experiments, they are rarely on the exact question of legislative interest. But GAO's methods are almost always more timely than those of a new randomized experiment and often of reasonable accuracy.

Second, randomized experiments provide a precise answer about whether a treatment worked (Cronbach et al., 1980). But the need for great precision may be low in many cases. For example, when much high-quality prior information exists about the treatment, a review of existing literature may be a better use of resources than would be a new randomized trial. When a causal question is of secondary interest to a noncausal question, such as whether services are being provided as intended, program monitoring procedures may be better. When an effect is so large and dramatic that no one doubts it resulted from the treatment, as with the dramatic effects of screening for PKU on PKU-based retardation among children, investing in an additional randomized experiment may be superfluous.

Third, randomized experiments can rarely be designed to answer certain kinds of questions. It is not possible to assign persons at random to variables that cannot be manipulated, such as age or race, or to manipulate events that occurred in the past, such as the effects of the death of President John F. Kennedy or of the Great Depression in the 1930s. It is unethical to assign persons at random to many manipulable events that cause significant harm, such as to cigarette smoking or to having a spinal cord injury.

Fourth, before conducting an experiment, a good deal of preliminary conceptual or empirical work must be done. The Federal Judicial Center (1981) recommends that, before an experiment is conducted, it should be demonstrated that the present conditions need improvement, that the proposed improvement is of unclear value, that only an experiment could provide the necessary data to clarify the question, that the results of the experiment would be used to change the practice or policy, and that the rights of individuals would be protected in the experiment. Similarly, the National Cancer Institute's five-phase model of testing a potential cancer control method suggests that, before a randomized experiment is conducted, the existing scientific literature should be identified and synthesized to see if an empirically supportable and testable hypothesis can be generated; pilot tests should be done to investigate the feasibility or acceptability of an intervention; studies assessing participation and adherence in the population should be conducted; data collection forms should be developed and validated; and quasi-experimentally controlled studies should be used to provide preliminary evidence about treatment effects (Greenwald & Cullen, 1984). Premature experimentation can be a great waste of resources—indeed, it can undermine potentially promising interventions for which there has not yet been time to develop recruitment procedures, identify and fix implementation problems, and serve the clientele long enough to make a difference.

DISCUSSION

The randomized experiment is often the preferred method for obtaining a precise and statistically unbiased estimate of the effects of an intervention. It involves fewer assumptions than other methods, the validity of those assumptions is usually easier to check against the data and the procedures used, and it requires less prior knowledge about such matters as selection processes and unit characteristics than do quasi-experiments, causal modeling, and selection bias models. Given all these strengths, it is easy to forget the many practical problems that can arise in implementing randomized experiments.

One practical problem concerns the feasibility and desirability of experimenting in particular cases. Some experimental manipulations are not ethical, as in the case of a physician deciding that a certain class of patients must be given a certain treatment and so cannot be randomized, or of an experimental treatment producing positive or negative effects that are so large that it would be unethical to continue to study them. Other times, it is not acceptable to wait the years that a well-designed and implemented experiment can take. Still other times, legal problems arise, not only because ethical violations can become legal problems but also because the law is often involved in certain experimental situations, for instance, when it mandates experimental evaluations of a program; when participants are directly under legal scrutiny, as with prisoners; or when legal systems are themselves the target of study.

A second practical problem is that a sufficiently large number of people (units) may not exist who are both eligible and willing to receive the treatment if assigned to it at random. Many is the experiment that has failed on this count. Frequently, especially with researchers who have never run a large field experiment before, the number of eligible people is vastly overestimated, as is the ease with which they can be located. When they are located, they often refuse to participate. In the worst case, the result is the death of the experiment for lack of participants.

A third practical problem is that the randomization procedure is not always properly designed and implemented. Sometimes this problem occurs because the researcher does not understand what random assignment is and so substitutes a seemingly haphazard assignment procedure. Or the researcher may introduce ad hoc adjustments to a random assignment procedure that seems to be yielding groups that are unequal before treatment, all the while thinking that these procedures are random when they are not. Other times the researcher correctly designs random assignment procedures but fails to create or supervise the procedures for implementing random assignment, so the assignment is implemented improperly. Whenever randomization is incorrectly or incompletely implemented, its benefits may be thwarted.

A fourth practical problem is that the treatment assigned is not always the treatment received. Participants may fail to fully receive the treatment to which they are assigned or may not receive it at all, as in the case of patients assigned to drug therapy who fail to take the drug or take only part of it. They may cross over to another condition (in a design that does not call for a crossover), as in the case of participants in a control condition who reapply for treatment and are accepted. Diffusion of treatment may occur through such means as treatment-related communication between participants in different conditions. Here, too, the participant is now receiving some part of both conditions. In all these cases, the intended treatment contrast is thwarted. If so, although the inference that *assignment to condition caused outcome* is still clear, the construct validity of the treatment is not clear. Hence it can be useful to prevent these failures of treatment implementation or measure their occurrence in many experiments in which pure treatment contrasts are desired.

A fifth problem is attrition. The randomized experiment does not just aim to make groups equivalent before treatment begins; it also aims to make groups equivalent at posttest in all respects except for differences in treatment conditions. Differential attrition from conditions after initial random assignment can vitiate this latter aim. Such attrition occurs often in field experiments. So preventing attrition, coping with attrition, measuring attrition, and analyzing data with attrition all become crucial adjunct topics to the study of the randomized experiment.

This chapter, being mostly about the design and logic of randomized experiments, has skirted all these problems in the interests of presenting the simplest case and its variants. But the researcher needs to know about these problems because they bear on the decision whether to use the randomized experiment at all, and if the decision is to do so, then they bear on how well the experiment is implemented and subsequently interpreted. So we turn to these problems in more detail in the next two chapters.