9

Practical Problems 1: Ethics, Participant Recruitment, and Random Assignment

Prac·ti·cal (prăk′tĭ-kəl): [Middle English *practicale*, from Medieval Latin *practicalis*, from Late Latin *practicus*, from Greek *praktikos*, from *prassein*, to make, do.] adj. 1. Of, relating to, governed by, or acquired through practice or action, rather than theory, speculation, or ideals: *gained practical experience of sailing as a deck hand.* 2. Manifested in or involving practice: *practical applications of calculus.* 3. Actually engaged in a specified occupation or a certain kind of work; practicing. 4. Capable of being used or put into effect; useful: *practical knowledge of Japanese.*

Eth·ic (ĕth′ĭk): [Middle English *ethik*, from Old French *ethique* (from Late Latin *ethica*) (from Greek *ethika, ethics*) and from Latin *ethice* (from Greek *ethike*) both from Greek *ethikos, ethical*, from *ethos, character.*] n. 1. a. A set of principles of right conduct. 1b. A theory or a system of moral values: "*An ethic of service is at war with a craving for gain*" (Gregg Easterbrook). 2. ethics. (*used with a sing. verb*) The study of the general nature of morals and of the specific moral choices to be made by a person; moral philosophy. 3. ethics. (*used with a sing. or pl. verb*) The rules or standards governing the conduct of a person or the members of a profession: medical ethics.

Re·cruit (rĭ-kroot′): [French *recruter*, from obsolete *recrute, recruit*, variant of *recrue*, from feminine past participle of *recroître, to grow again* from Old French *recroistre:* re-, *re-* + croistre, *to grow* (from Latin *crescere*).] v. *tr.* 1. To engage (persons) for military service. 2. To strengthen or raise (an armed force) by enlistment. 3. To supply with new members or employees. 4. To enroll or seek to enroll: colleges recruiting minority students. 5. To replenish. 6. To renew or restore the health, vitality, or intensity of. **re·cruiter** *n.* **re·cruitment** *n.*

EVEN VERY good experiments encounter practical problems. For instance, the Perry Preschool Program experiment violated randomization protocols in minor ways, and even their heroic effort to follow participants until age 27 was only 95% successful. Such problems are sometimes so severe that they thwart the experiment completely. In the Madison and Racine Quality Employment experiment, the pool of applicants was too small, and the treatment program could not develop many good jobs for those who did apply. Consequently, the intervention could never show the expected impact on getting good jobs, so the experiment was ended prematurely (Greenberg & Shroder, 1997).

Experienced researchers know that designing a good experiment is only half (or less!) of the battle. Successfully implementing that experiment requires coping with many practical problems. In this chapter and the next, we outline these problems and describe strategies to address them. This chapter focuses on problems that occur early in experiments: ethical and legal issues, getting enough participants to be in the study, and correctly implementing random assignment. In the next chapter, we address problems that occur later: treatment implementation issues and coping with post-assignment attrition. Except for issues that are specific to random assignment itself, most of these practical problems apply to nonrandomized experiments just as much as to randomized ones, and some of them (e.g., tracking participants, using appropriate informed consent, guarding confidentiality) apply to nonexperimental research, such as surveys, as well.

ETHICAL AND LEGAL ISSUES WITH EXPERIMENTS[1]

Ethics should be considered from the very start of the process of designing an experiment. Here we address a few key ethical and legal questions that philosophers, lawyers, and scientists have raised about both experimentation and randomization:

- the ethics of experimentation on human beings;
- the ethics of withholding a potentially effective treatment from control or comparison participants;
- the ethics of random assignment compared with alternatives such as assignment based on need;
- the conditions under which experiments might be discontinued for ethical reasons; and
- some legal problems that bear on experiments.

1. This section focuses on problems specific to experimentation. However, the conduct of science involves many other ethical problems. For example, concerns about fraud in some clinical trials (e.g., Ranstam et al., 2000) have led to useful protocols for ensuring the integrity of data (Knatterud et al., 1998); similarly, the management of data to insure its integrity presents a host of practical problems with ethical features (McFadden, 1998).

The Ethics of Experimentation

In the name of scientific experimentation, great wrongs have been done. Abusive medical experiments during World War II, especially in Nazi concentration camps, are well-known (Greenberg & Folger, 1988). In the United States, researchers in the Tuskegee syphilis study withheld effective treatment from African-American males suffering from syphilis so that scientists could observe the long-term course of the disease (J. Jones, 1981). Less extreme examples are prevalent (Beecher, 1966; Veatch & Sollitto, 1973). Indeed, it seems inevitable that a methodology that depends on manipulation would encounter objections to such manipulations—especially when the manipulation can cause harm, as with medical treatments or decisions about punishment of criminals. To counter these problems, experimenters use three sources of help:

- Ethical Codes
- Informed Consent
- Institutional Review Boards

Ethical Codes and Principles

To reduce abuse and foster ethics, governments have adopted various codes of ethics for scientific research, for example, the Nuremberg Code adopted by the United Nations General Assembly (Nuremberg Code, 1949; although Miké, 1990, points out that reasonable ethical codes were present in Nazi Germany and did not prevent abuse); and ethicists have proposed other similar ethical standards (Emanuel, Wendler, & Grady, 2000; World Medical Association, 2000). The U.S. Public Health Service's Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979)[2] proposed three ethical principles for scientific research with human participants:

1. *Respect for Persons:* that individuals are autonomous agents with the right to decide whether to enter a study (hence the need for informed consent) and, if their autonomy is diminished by a disability, have the right to be protected from harm;
2. *Beneficence:* that researchers must maximize benefits and minimize harms to participants (hence the need to reveal potential harms and benefits);
3. *Justice:* that the benefits and harms of treatment be distributed fairly (hence the need to recruit participants fairly), and that persons not be deprived of efficacious treatments to which they would otherwise be entitled (hence the need to inform of alternative treatments).

The latter requirement was motivated partly by the observation that abuses in experiments often fell disproportionately on disadvantaged or vulnerable persons

2. See http://grants.nih.gov/grants/oprr/humansubjects/guidance/belmont.htm.

such as Black males in the pre–civil rights era in the Tuskegee study or concentration camp prisoners during World War II.

Informed Consent and Experiments

To operationalize these principles, participants are often asked to give their written *informed consent* to being in the experiment (Protection of Human Subjects, 1983[3]). The U.S. Public Health Service requires that human participants in research studies that it funds should read and sign a consent statement that includes:

1. A statement that the study involves research, an explanation of the purposes of the research, the expected duration of the participant's participation, a description of the procedures to be followed, and identification of any procedures that are experimental.
2. A description of any reasonably foreseeable risks or discomforts to the participant.
3. A description of any benefits to the participant or to others that may reasonably be expected from the research.
4. A disclosure of appropriate alternative procedures or courses of treatment, if any, that might be advantageous to the participant.
5. A statement describing the extent, if any, to which confidentiality of records identifying the participant will be maintained.
6. For research involving more than minimal risk, an explanation as to whether any compensation and any medical treatments are available if injury occurs and, if so, what they consist of or where further information may be obtained.
7. An explanation of whom to contact for answers to pertinent questions about the research and research participants' rights and whom to contact in the event of a research-related injury to the participant.
8. A statement that participation is voluntary, that refusal to participate will involve no penalty or loss of benefits to which the participant is otherwise entitled, and that the participant may discontinue participation at any time without penalty or loss of benefits to which the participant is otherwise entitled.

Boruch (1997, pp. 44–49) provides sample informed consent forms from past experiments. New experimenters should consult those with experience to help generate a contextually appropriate informed-consent protocol.

The Public Health Service (PHS) does not require informed consent for surveys or confidential educational tests. For prisoners, pregnant women, or those with diminished mental capacity, more stringent requirements apply (Federal Judicial Center, 1981; Mastroianni, Faden, & Federman, 1994; Stanley & Sieber, 1992). Research with children requires either active or passive consent from parents (Esbensen et al., 1996). Some other federal agencies in the U.S. government

3. See http://ohrp.osophs.dhhs.gov/humansubjects/guidance/45cfr46.htm.

have adopted these rules, including the Departments of Education, Agriculture, Justice, and Health and Human Services (Boruch, 1997), as have many professional associations whose members study human participants. For example, the American Psychological Association's (1992) ethical principles require many of these procedures of psychologists who study human participants (J. Greenberg & Folger, 1988; Sales & Folkman, 2000; Sieber, 1992). Many individual researchers and research firms use these procedures voluntarily (e.g., Gueron, 1999; Orr, 1999), both for ethical reasons and because obtaining informed consent can help protect the experimenter against liability for harms that experiments might cause.

Institutional Review Boards

The Code of Federal Regulations (Protection of Human Subjects, 1983) also established a limited requirement for Institutional Review Boards (IRBs) at institutions receiving PHS funding for research with human participants, including many universities, government agencies, and private research firms. The IRB monitors research with human participants by reviewing the experimental and informed consent procedures for ethical problems. It may also review the scientific quality of research, such as whether the experiment has adequate statistical power, because to conduct a seriously underpowered experiment would waste both resources and the time of the participants.

These particular procedures are not always uniformly applicable to all human research. They were developed mostly in medical research, and other substantive areas sometimes have different needs. For example, the Guiding Principles of the American Evaluation Association (American Evaluation Association, 1995) do not include the Belmont Report's (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979) principle of beneficence because evaluations by their very nature sometimes do harm to the interests of some people—for example, identifying poorly performing programs is often part of the evaluator's contract, even though the program might be terminated and program personnel lose their jobs (Shadish, Newman, Scheirer, & Wye, 1995). Similarly, some researchers object that telling research participants about other services to which they are entitled without being in the research might destroy our ability to form control groups (Rossi, 1995); and the American Psychological Association allows deception under limited conditions. But in matters of informed consent, our belief is that exceptions to customary practice require careful thought and consultation with experts and colleagues about the justification.

Withholding a Potentially Effective Treatment

Arguments for and Against Withholding Treatment

In experiments, treatments are deliberately withheld from or reduced for some participants, or a different (and potentially less effective) treatment is given to

some participants than to others. This practice can be ethically questionable if beneficial treatments are withheld from persons who might need or deserve them. Fetterman (1982), for example, questioned the ethics of assigning disadvantaged high school dropouts to a no-treatment control when they had expressed interest in a remedial "second-chance" treatment because this second chance might be their last chance. Similarly, a debate in AIDS research concerns whether treatments that are promising in preliminary research should be released to patients before results from more extensive tests are done (Marshall, 1989). For both ethical and practical reasons, there is some obligation to err on the side of benefiting the patient when doubt exists about the most ethical action. Consequently, the health care establishment has experimented with procedures such as the establishment of "group C" cancer therapies that could be given to patients prior to typical rigorous FDA testing; the Fast Track Accelerated Approval process for moving drugs into use more quickly; and the "parallel track" structure for distributing AIDS drugs that are safe and promising in preliminary tests (Marshall, 1989; Expanded Availability, 1990). AIDS research has benefited enormously from interaction with the AIDS patient community and their advocates. Many of their concerns have now been better addressed, both within and outside the accepted principles of experimental design (Ellenberg, Finkelstein, & Schoenfeld, 1992).

However, good ethical arguments exist *for* withholding treatments in experimental comparisons. If such treatments are released too early, toxic side effects may not be found until after they have harmed people; long-term effects are more difficult to study; drug companies may sell treatments that have little effect; and physicians who prescribe such drugs may become liable for toxic effects (and should follow special informed consent procedures). The tradeoffs are obvious, but the solutions are not.

Withholding treatment is sometimes ethical when scarce resources make it impossible to provide treatment to everyone. In the former case, the investigator can use a crossover design (Fleiss, 1986; Pocock, 1983) to provide the treatment, or a refinement of it, at a later date to control participants if the treatment is successful and resources allow. However, this solution works best when the problem being treated is not progressive and does no permanent damage. For example, a patient with cancer may experience a more rapid progression while being treated with a less effective comparison treatment, and the later crossover to treatment may not compensate for that loss (Marquis, 1983). Similarly, a disadvantaged high school dropout with few options who is denied entry into a second-chance educational program may experience lifelong educational, social, and economic damage if no other chance comes along (Fetterman, 1982). But if the problem will not get worse without treatment, this option helps ensure that everyone will finally receive a treatment that turns out to be useful. It also benefits the experimenter, who can confirm preliminary findings on the former comparison group participants.

Withholding treatment can also be ethical when alternative treatments, each of approximately equal desirability, are compared. The strategy of "planned vari-

ations" in the evaluation of educational programs is an example (e.g., Rivlin & Timpane, 1975). However, the strategy is difficult to implement, as researchers often found more variation within a planned treatment over sites than they found between different treatment variants. Further, different variations are often targeted at different parts of the same problem (e.g., cognitive performance versus academic self-concept) and so should include measures of those particular problems, as well as measures of the general problem to which all the variants are aimed (e.g., educational achievement). Another example is a study of computer-assisted instruction (CAI; Atkinson, 1968). Entire classrooms were randomly assigned to receive CAI in either mathematics or English, and each group was tested on both subjects, so that mathematics was the experimental and English the control topic for some students, whereas the reverse was true for other children. This design works when several equally pressing problems are to be treated and when the particular form of one treatment (CAI in English) would not be expected to affect very much the tests pertaining to the other form of the treatment (CAI in mathematics). Such conditions do not hold in many areas, such as AIDS research.

Options When Withholding Treatment Is Problematic

Most other cases of withholding treatment are problematic, especially for severe, deteriorating, or permanently damaging problems. Then, useful options include:

- Using dose-response designs
- Offering all participants an intervention prior to randomization
- Using a "treatment-on-demand" control

First, in dose-response studies participants receive one dose from a range of doses from lower to higher. For example, in a spousal assault intervention experiment, those arrested could be assigned either to normal arrest and release condition in which they were held only a few hours or to an extended arrest condition in which they were held as long as the law allowed (Boruch, 1997). Once the dose of the intervention is reduced to a certain point, the lower end is like a placebo control (although placebos can themselves show a dose-response relationship; Clark & Leaverton, 1994).

Second, one can offer all participants an intervention that occurs *prior* to randomization so that everyone gets something. Some sites used this option as part of the Rockefeller Foundation's Minority Female Single Parent (MFSP) program (Boruch, Dennis, & Carter-Greer, 1988) that provided employment training, day care, and other support services to economically disadvantaged minority women who were single parents. Women were randomly assigned either to receive the intervention or to a control group that could not receive some of the employment and training services. Before random assignment, one site held meetings to help all women to develop a plan to solve some of their problems in case

they were assigned to the control condition. Of course, this change may reduce the chance of finding a treatment effect.

A third solution is the treatment-on-demand (TOD) control condition, particularly if fewer participants are assigned to that control than to treatment. Snyder and Wills (1989), in a study of marital therapy, created such a control condition to which fewer couples were assigned than to the treatment conditions. Couples in this control condition could request therapeutic consultations up to 1 hour biweekly for crises that jeopardized the couple's ability to tolerate the 3-month waiting period. Couples who requested more than three consultations were dropped from the study. This procedure has fewer ethical problems than one with a control group receiving no treatment. Attrition is not severe if few couples actually reached the attrition threshold or if such couples are kept in the control anyway and then analyzed using methods discussed in the next chapter.

Imposing Undesirable Treatments

Finally, some experiments involve imposing a potentially undesirable treatment. Such treatments may contain a noxious element, such as a nausea-producing agent in certain aversion-training procedures. They may cause harmful side effects, such as the loss of hair and fatigue in chemotherapy. They may be costly or time-consuming. For example, one study randomly assigned 21 schools to two treatment conditions and one control condition (Moberg, Piper, Wu, & Serlin, 1993). The intensive-treatment condition required a large investment of time and resources by the schools, so some schools refused to be in that condition. This problem was caught early and remedied by allowing schools to self-select their preferred treatment condition with the understanding that they would then be assigned at random either to that treatment or to a control but not to the treatment they objected to. Solutions discussed in the previous sections might apply here to these kinds of experiments, as well.

The Ethics of Random Assignment

Many experiments involve the distribution of scarce and desired resources, such as income supplements, educational grants, or new medicines. Decisions about who will receive these resources must be made ethically. Randomized experiments make the decision by lottery, no matter what the applicants' needs or merits, no matter whether they applied first or last, and no matter who they know or what their connections or power may be (cronyism). We have seen no arguments for treatment allocation by cronyism, for this seems to violate our basic sense of fairness. We could imagine an argument for allocation by order of application, for example, to reward those who made the effort to apply first. But the argument would be weak because effort is confounded with the resources one has to apply—for example, those who apply first may have better access to transporta-

tion or more leisure time. So need, merit, and randomization are usually the main options.

The case for assignment by need is strongest when a treatment is known to be the most effective in meeting that need. For example, antibiotics are widely acknowledged to be the most effective treatment for pneumonia, so it is difficult to imagine a case in which it would be ethical to withhold them from such a patient. A similar case can be made for assignment by need to a proven compensatory education program intended to help students from impoverished inner-city schools. Examples of assignment by merit are similarly easy to find, as with the awarding of a National Merit Scholarship for high performance on the National Merit Exam. Bolstering these arguments is the fact that the regression discontinuity design (Chapter 7) can yield unbiased estimates of effects when assignment is made by need or merit (although it is less powerful than the randomized experiment and has its own implementation problems).

Arguments Against Randomization

Given the arguments for assignment by need or merit, some philosophers assert that randomization is ethical only if the conditions being compared may be therapeutically equivalent and if no better treatment exists (Marquis, 1983). They argue that even if the difference between two treatments is based on poor designs and is small, say, only one chance in one thousand of doing better with one treatment, the ethical principle of autonomy contained in the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979) requires that participants be told of these results in the informed-consent process so that they can judge the relevance and weight of such evidence for themselves—especially in areas such as medicine, in which participants usually have some discretion in choosing treatment. These scholars ask the reader to imagine being a cancer patient whose life is at stake before assuming that one is not entitled to all information, no matter how preliminary. Failure to recognize such simple realities can have drastic consequences for the feasibility of the research, as in the example provided by Berg and Vickrey (1994) in which "funding was withdrawn for another trial of epilepsy surgery after an insufficient number of patients were willing to have the decision of whether to have part of their brains surgically removed decided at random" (p. 758).

Other arguments against randomization appeal to practical realities that impede it. Those who sign informed consents may frequently not understand either the forms or their legal rights (T. Mann, 1994). Attitudes toward randomization are not always favorable, even when resources are scarce (Hillis & Wortman, 1976), especially if harm may occur from some treatment conditions (Innes, 1979). Women in the Rockefeller MFSP program thought randomization was less fair than assignment by need or by test scores (but more fair than assignment by "first come, first served" or by "who you know"). About 25% of patients eligible for the NIMH Collaborative Depression Project refused to sign the informed

consent because it included a consent to randomization (Collins & Elkin, 1985). Some participants wrongly believe that by signing the form, they have signed away their rights to sue for negligence (T. Mann, 1994). Service providers can react equally negatively, sometimes blaming randomization for problems actually caused by other factors, such as poor referral practices (Lam et al., 1994). Respondents may not believe that a truly random procedure was used, particularly if they receive a less desirable treatment (Boruch, Dennis, & Carter-Greer, 1988; Lam et al., 1994; Wortman & Rabinowitz, 1979), though the latter can sometimes be ameliorated by doing random assignments in a forum open to public participation and scrutiny, as with the U.S. military draft of the 1970s.

Arguments for Randomization

Advocates of randomization respond that we frequently do not know which treatment works best, particularly for innovative treatments with little empirical track record (Miké, 1989). Especially when the sample size is small, as it is likely to be in small trials, or at the time interim results first appear, sampling error is quite likely and quite large, so early results can be misleading. Often the data available about such innovations are based on poorly controlled or executed research that has been advanced to the stage of a proposed clinical trial precisely in order to get larger samples, better designs, and more trustworthy outcome data. In order to advance as far as a clinical trial, the innovation has to show promise, but that promise often proves to be illusory, or worse yet, the innovation causes harm. For example, in observational studies beta-carotene reduced cancer risks, but in clinical trials it increased risk (Liebman, 1996). Indeed, a review of randomized trials showed that medical innovations produce better results than standard treatment only about half the time (Gilbert, McPeek, & Mosteller, 1977b). As Chalmers (1968) put it: "One has only to review the graveyard of discarded therapies to discover how many patients might have benefited from being randomly assigned to a control group" (p. 910). Given the complexities of the statistical issues surrounding the uncertainty of evidence, the researcher presumably has some obligation to point to better and worse evidence.

Other Partial Remedies

Clearly this controversy has no quick fixes. Following are additional solutions that are sometimes used (see Table 9.1). Random assignment can include a "safety valve" mechanism, holding a certain number of program slots for the most needy applicants who are not entered into the experiment, though this may decrease power if the most needy are also most likely to benefit. This strategy was used in the Rockefeller MFSP program, in which staff could fill 10% of program slots with women they judged to be most needy (Boruch et al., 1988).[4] One can divide participants into strata from least to most needy and assign proportionately more of the most

4. If the judgment can be quantified, this becomes a regression discontinuity design.

TABLE 9.1 Partial Remedies to Ethical Problems with Random Assignment

1. Use a regression discontinuity design to assign based on need or merit instead of randomly.
2. Use a safety valve mechanism for treating the most needy participants.
3. Divide participants into strata by need, assigning proportionately more of the most needy to treatment.
4. Assign proportionately more participants to the treatment in highest demand.
5. Use a dose-response design.
6. Use an adaptive design to increase the proportion of assignment over time to the most successful condition.
7. Use the informed-consent procedure to ensure that participants are willing to be assigned to all conditions.
8. Use public lotteries for assignment to increase perceived fairness.

needy persons to treatment (Coyle et al., 1991); or similarly, one can assign participants to conditions that vary only in the intensity of treatment (say, dose), so that no participant goes without treatment entirely. When the experiment can use multiple sites but has only limited resources, the treatment can be implemented at some of the sites. A final option is to use adaptive designs that use Bayesian logic in which the proportion assigned to a condition increases over the life of an experiment as the number of successes produced by that condition increases (Flournoy & Rosenberger, 1995; Kadane, 1996; Tamura, Faries, Andersen, & Heiligenstein, 1994).

Discontinuing Experiments for Ethical Reasons

Ethics sometimes demands that ongoing experiments be halted if negative side effects unexpectedly occur or if one treatment condition is producing dramatically better results than another (Marquis, 1983). This happened in the Physicians' Aspirin Study, which examined whether aspirin reduced heart attacks in physicians more effectively than a placebo (Steering Committee of the Physicians' Health Study Research Group, 1988). Of 22,071 patients, 104 had heart attacks while on aspirin and 189 had heart attacks while on the placebo—an effect that objectively seemed small statistically ($r = .034$)[5] but that was large enough to end the study early on grounds that it was unethical to withhold such an effective treatment from controls. Similar decisions to end randomized trials on ethical grounds have occurred in other areas of medicine (Marx, 1989; R. Rosenthal, 1994), though we know of no examples in the social sciences.

Hence it is common to do preliminary analyses of early returns at fixed intervals to determine whether to stop the experiment (Choi & Pepple, 1989; Choi, Smith, & Becker, 1985; S. Emerson, 1996; Spiegelhalter, Freedman, & Blackburn,

5. This correlation is somewhat higher ($r = .16$) if computed differently (Haddock, Rindskopf, & Shadish, 1998); in either case, however, the effect size is undeniably small.

1986; Tan & Xiong, 1996). The issue is not entirely statistical, for it also involves determining who decides to stop the study and whether treatment providers get to see the preliminary results if the experiment is not stopped. These decisions are often made by a data and safety monitoring board (Dixon & Lagakos, 2000). In addition, important outcomes may be revealed only after long periods of time, so terminating a trial after only short-term outcomes have been observed may miss crucial findings (Armitage, 1999). Presenting interim results without ensuring that these problems are clearly understood may mislead participants into selecting a treatment that later proves ineffective or harmful.

Legal Problems in Experiments

Ethical problems can be grounds for legal action. This occurred in the New Jersey Negative Income Tax Experiment. Confidentiality was promised to participants. But a Mercer County grand jury, suspecting that some respondents were fraudulently receiving both experimental payments and welfare, subpoenaed the experimenters' records in order to identify respondents. This action placed the experimenters in a difficult dilemma: to protect possible lawbreakers who had been promised that their responses would be kept confidential, the experimenters would have had to defy the grand jury. Eventually a legal settlement was made by the experimenters out of court.

The law provides certain legal guarantees of confidentiality, such as prohibitions against revealing Internal Revenue Service (IRS) income records of individuals in most circumstances or the confidentiality of the physician-patient relationship. However, the confidentiality of research data is not routinely guaranteed in the law. No matter what an informed consent says, researchers may have to respond to a court subpoena, a warrant for their data, or a summons to appear in court with their data to answer questions. Research participants can also be summoned to court to answer the same questions that they were asked in an interview—although these are rare occurrences in practice (Cecil & Boruch, 1988). Some laws have been enacted to safeguard the confidentiality of data in specific cases. Boruch (1997) lists relevant statutes and the kinds of guarantees given—for example, immunity from legislative inquiry, provisions for secondary analysis of data, and so forth. In research in which legal complications are possible, such as with drug abusers or parole violators, the researcher should identify statutes that are pertinent to such research. However, many of these statutes have never been tested in court, so the extent to which they can be relied on is unclear. Hence researchers should consider using research procedures that can ensure confidentiality, such as using randomized response methods or determining not to gather any data about potential identifiers, and, just as important, they should ensure that such procedures are followed.

The legality of randomization has sometimes been challenged; for example, in criminal justice research some have argued that randomization is arbitrary

and so violates both our sense of fairness and the due process and equal protection rights of the accused (Baker & Rodriguez, 1979; Erez, 1986; but see Lind, 1985). This complaint is also heard in entitlement programs such as social security or welfare. Resolutions are legislative, judicial, and procedural (Boruch, 1982, 1997). Sometimes, specific laws authorize randomization, as in the evaluation of the Job Opportunities and Basic Skills Training Program, for which a federal law mandated that "a demonstration project conducted under this subparagraph shall use experimental and control groups that are composed of a random sample of participants in the program" (Family Support Act, 1988). At other times that randomization has been contested in courts, the courts have ruled that it is legal, establishing some case law precedent for later uses of similar randomization (Breger, 1983). Sometimes procedures can be approved to authorize randomization after a law is passed. Gueron (1999) describes a case in which a welfare reform was mandated by law. Although participants could ordinarily not opt out of these legally imposed program requirements, it was possible to arrange afterward for some recipients to be excused from them in the interests of the evaluation. Informed-consent procedures provide some defense, especially if an IRB reviewed and approved the work. Still, many managers and service providers fear being sued if they participate in the experiment. In one experiment on police methods for dealing with domestic violence, the city government initially demanded that the experimenters provide liability insurance coverage for lawsuits incurred as a result of their participating in experiments (Dunford, 1990). Fortunately, this case was successfully resolved without the purchase of insurance.

A report by the Federal Judicial Center Advisory Committee on Experimentation in the Law (Federal Judicial Center, 1981) noted that the law has long viewed experimentation as a legitimate means of exploring public policy issues but with many stipulations. The principle of equality of treatment requires that individuals who are similar in relevant ways be treated similarly. Randomization can violate this principle if needy people are randomly divided into conditions rather than being treated the same way. Regarding benefits and harms, they noted that large harm to individuals is typically not justified in the law by appeals to some larger benefit to those who might receive the treatment in the future. The committee also expressed concern about any experiments that might undermine public faith in the justice system. The committee discussed how these principles might apply to harm caused by disparate treatments, the use of mandatory treatment, confidentiality issues, and deception in experiments. The committee recommended that before an experiment be conducted, it should be demonstrated that:

- the present conditions need improvement,
- the proposed improvement is of unclear value,
- only an experiment could provide the necessary data to clarify the question,
- the results of the experiment would be used to change the practice or policy, and
- the rights of individuals would be protected in the experiment.

RECRUITING PARTICIPANTS TO BE IN THE EXPERIMENT

An experiment cannot begin without locating participants who are willing and eligible to participate. Failure to do so can kill an experiment entirely, as with the Madison and Racine Quality Employment experiment cited at the start of this chapter, or can reduce its power dramatically. Moreover, the characteristics of those who do eventually participate will affect both construct and external validity in obvious ways. For example, the Systolic Hypertension in the Elderly Program experiment (Cosgrove et al., 1999) sent out over 3.4 million letters of invitation, contacted 447,921 screenees, found 11,919 eligible people, and randomized 4,736 to conditions—that is, those who were assigned were just a small subset of those eligible, who were themselves only a small subset of volunteers from a large population that also included nonvolunteers who might have been eligible if screened. In one alcohol treatment outcome study, 76% of a pool of alcoholics were screened out, another 15% refused random assignment to conditions, and the remaining 9% differed greatly from the original pool on many variables (Howard, Cox, & Saunders, 1988). And Pearlman, Zweben, and Li (1989) showed that participants who were solicited for alcohol treatment research by media advertisements differed significantly from the usual alcohol clinic patients on many key characteristics.

Sometimes the problem is that the target population is not clearly defined but merely labeled as unemployed, inner-city residents, or educationally disadvantaged children. More often, the target population is defined but cannot be located (Bickman, 1985; Boruch, 1997; Boruch et al., 1988; Dennis, 1988; Dunford, 1990; Ellenberg, 1994; Orwin, Cordray, & Huebner, 1994). To ensure that they can be located, experimenters can use: (1) preexperimental surveys to locate and characterize potential participants, (2) pipeline studies to follow what happens to them over time, (3) pilot tests of the solicitation procedures to see who will learn about the experiment and who will attend if eligible, (4) trained outreach specialists skilled in communicating and generating interest about a program, (5) intake specialists who aggressively recruit potential participants, and (6) efforts to learn what features of the intervention cause people to decline to enroll, such as lack of child care or finding the training to be in a career that they do not ultimately want to pursue.

If the number of eligible participants is lower than anticipated despite these efforts, the researcher can (1) extend the time frame for the experiment if time and resources are available and if program staff are willing; (2) divert additional resources to intensified outreach efforts; (3) alter eligibility requirements so that more participants are eligible, though this may require testing interactions between eligibility requirements and treatment; (4) reduce the proportion assigned to treatment if power analyses suggest it would still be feasible to find an effect; or (5) terminate the experiment, which is sometimes better than spending funds

on an experiment that cannot find the expected treatment effect, especially if the decision is made early in the life of the experiment.

From among those selected as the target sample of interest, the volunteers who agree to be in the experiment may have different characteristics than nonvolunteers do. For example, Klesges et al. (1988) offered a workplace smoking cessation program to 66 eligible smokers in two work sites. The 44 who agreed to participate in the experiment differed from those who did not participate in being less successful in previous attempts to quit, in having smoked longer, and in perceiving themselves to be more vulnerable to smoking-related diseases. Such characteristics suggest that these 44 participants were more tenacious smokers who would presumably find it more difficult to quit than those who did not volunteer, possibly biasing the experiment toward finding smaller effects. Notice how this situation affects many kinds of validity—power is reduced given the smaller sample size, the program can only be described as effective with volunteers, and we do not know if the program would work with nonvolunteers.

Sometimes decisions about who is eligible for the experiment are affected by awareness of randomization procedures (Dunford, 1990). For example, in a randomized experiment testing three different methods of police response to domestic violence, individual police officers knew the condition to which the next case would be assigned. They also had responsibility to judge if a case met experimental eligibility criteria. As a result, if they did not believe the next case merited the condition to which it had been assigned, they sometimes defined the case as ineligible. If this problem occurred systematically for one condition more than the others (e.g., if officers were reluctant to assign comparatively minor disputes to the arrest condition and so defined those cases as ineligible when the arrest condition was next to be assigned), both internal and external validity would be affected simultaneously. Here, separating the eligibility and recruitment judgment from the assignment process would help.

Some researchers have proposed design and statistical methods for estimating the effects of these preassignment selection problems. Marcus's (1997a) method requires conducting parallel randomized and nonrandomized experiments, both experiments being identical except that participants willing to be randomized are in the randomized experiment but those wishing to choose their treatment are in the nonrandomized experiment. Ellenberg's (1994) method studies a random sample of participants at each stage of elimination (e.g., those with a diagnosis who are eliminated as inappropriate for treatment, those appropriate patients who cannot be screened, those who are screened but deemed ineligible for randomization, and those who are randomized but refuse to accept assignment) to provide data about the biases that are introduced at each step. Braver and Smith's (1996) design randomizes the entire pool of eligible participants to three conditions: (1) a lottery condition in which participants are randomized to both treatment and control, (2) an invitation to treatment in which participants are offered the chance to participate in treatment, and (3) an invitation to control in which participants are offered the chance to participate in the control. In all three conditions, some

participants will refuse participation; but the design takes advantage of these refusals by using a number of randomized and nonrandomized comparisons that shed light on what the effects of treatment might have been in the full population. This design might be usefully combined with the instrumental variable statistical methods presented in the next chapter to estimate the effects of receipt of treatment rather than just invitation to treatment.

IMPROVING THE RANDOM ASSIGNMENT PROCESS

After participants are recruited to a randomized experiment, they have to be randomly assigned to conditions. Successfully implementing and monitoring random assignment in field settings requires great practical expertise and craft knowledge. Few experimenters have that knowledge, in part because statistical textbooks rarely include it and in part because many researchers have little prior experience doing experiments (Haveman, 1987). More than half those funded to conduct an experiment by the National Institute of Justice over 15 years had never conducted one before (Dennis, 1988).

Methods of Randomization

Units can be randomly assigned to conditions in many ways (Table 9.2). We present here some common techniques and their advantages and disadvantages (see also Kalish & Begg, 1985; Lachin, Matts, & Wei, 1988). The techniques sometimes can be used jointly if a particular design has characteristics that would warrant it.

Simple Random Assignment

The procedures for implementing the best known random assignment procedures—the toss of a coin or the roll of a die—are so well known as to need no explanation. Similar possibilities include shuffled cards, spinners, roulette wheels, and urns filled with numbered balls. These procedures have considerable public relations value when the decision that random assignment makes is sensitive and public (as with winning a lottery or being drafted into the military) given the personal consequences of winning or losing. When implemented properly, these easily understood and publicly transparent procedures result in perfectly good random assignments.

But they have two key practical drawbacks. One stems from the physical structure of some devices. A coin only has two sides and so is best adapted to assignment of units to two conditions; a die has six sides, making it useful for up to six conditions. Yet many experiments, for example a 3×3 factorial design, have more conditions than this. Coins and dice can be adapted to more conditions, but

TABLE 9.2 Varieties of Random Assignment

1. Simple Random Assignment
   - Any procedure for assigning units to conditions by chance with nonzero probability (without replacement).
   - Typically done with table of random numbers or computer-generated random numbers.
2. Restricted Random Assignment to Force Equal Sample Sizes
   - Particularly useful with small sample sizes to prevent severely unequal splits over conditions.
   - Equal sample sizes tend to maximize power for testing treatment main effects in many (but not all) designs.
   - Preferred method is to assign from matches or strata (see #7 below).
3. Restricted Random Assignment to Force Unequal Sample Sizes
   - Can be done to cope with practical limitations, such as restriction in the number of units that can receive treatment or ethical objections to depriving many participants of treatment.
   - Can increase power to test certain limited hypotheses (see optimal design theory).
4. Batch Randomization
   - Refers to cases in which small groups with more units than experimental conditions, but not the whole sample, are available to be randomized to conditions.
5. Trickle Process Randomization
   - Refers to cases in which units trickle in slowly and assignment must be made from batches that are smaller than the number of conditions.
   - The key problem is to ensure that desired proportions are assigned to conditions over time.
6. Adaptive Randomization Strategies
   - Methods for correcting imbalances in the desired proportions assigned to conditions by changing the proportions over time.
   - Can be unbiased if unit characteristics remain stable over time; if not, certain analytic adjustments may be required.
7. Random Assignment from Matches or Strata
   - Placing units into groups and then assigning separately for each group.
   - In matching, groups contain as many units as conditions; in stratifying, groups contain more units than conditions.
   - Always helps control proportion assigned to conditions.
   - If matching or stratifying variable is related to outcome, this can also increase the power of the design.
   - Useful to match or stratify on variables expected to interact with treatment.

it can be logistically complex to do so, and the possibilities for errors increase and public transparency may decrease. The second drawback is that these procedures can be biased by systematic behavior by experimenters, as when they (perhaps unknowingly) always place the coin heads up prior to tossing it. For example, the 1970 draft lottery used capsules containing birthdays that were chosen blindly out of a 2-foot-deep bowl—seemingly random. But the capsules were initially placed

in a container in chronological order, and that container was probably not shuffled enough. When the capsules were poured from the container to the bowl with no further mixing, the early birth dates were on top of the bowl and were overrepresented among the first picks that sent some young people into the military (Fienberg, 1971). When these more publicly accepted procedures must be used, Fienberg (1971) suggests ways to increase the likelihood that the randomization process is reliable.

The main alternative is to use a table of random numbers. Many statistics texts contain them (e.g., Fleiss, 1986; Rosenthal & Rosnow, 1991). They are convenient and useful for experiments with any number of conditions, and they are often generated by computer algorithms that severely curtail the possibility of systematic biases. Over the years, they have been examined for many different kinds of latent nonrandomness, and biased columns or pages have been deleted and substituted. Though they are not perfect, such tables are the most accessible source of reliable random numbers.

Another alternative is to use a computer to generate one's own list of random numbers. Most common statistical packages can do this, including SPSS, SAS, and Excel (Appendix 9.1). These programs have great flexibility, but they can still have problems. For instance, one large city school system assigned eligible students at random to either a regular city school or to a new and innovative magnet school that spent four times more per pupil. A study of how the magnet school affected achievement and attitudes showed that attempts during one year to make the computerized random number generator more efficient actually made it more biased (Zigulich, 1977)! In the Spouse Assault Replication Program (SARP; Boruch, 1997), researchers discovered that a string of random numbers was being repetitively duplicated when the computer was rebooted after "going down" because the program started anew with the same algorithm generating the same numbers. Further, if the researcher does not fully understand the programming algorithms, a faulty randomization may result. Most researchers are well served if they continue to rely on tables of random numbers.[6]

The general principle in using any of these methods is to make all the allocation decisions as random as possible. Consider a psychotherapy experiment in which the experimenter wishes to assign 20 clients to either a treatment or control condition. First, the experimenter might pick a random start in a table, which may be done haphazardly by simply turning to a random page, closing one's eyes, and touching a spot on the page (Wilson, 1952), or more formally by using numbers picked haphazardly to identify the page, column, and row for the random start. Then the researcher simply moves down the column or across the row from

6. Creative researchers have invented many more randomization techniques. In the Spouse Assault Replication Program (Boruch, 1997; Garner, Fagen, & Maxwell, 1995), for example, the exact time of each complaining phone call to police was recorded to the second by a computer. All cases ending with an even-numbered second were assigned to the arrest condition for the alleged assault, and all those that ended in an odd-numbered second were assigned to the mediation condition.

that starting point, searching in this case for either the numbers 1 or 2 to indicate that the participant is assigned either to treatment or control (the pair can be 1 or 0, or any other plausible pair; or one could use odd numbers to assign participants to one condition and even numbers to the other). Suppose the numbers after the random start are 73856 20392 34948 12637. Moving left to right, the number 2 is encountered first, so the first participant is assigned to control. Continuing the search for 1 or 2, one encounters 2, 1, and 2 in that order. So the second and fourth participants are assigned to the control condition, but the third participant goes to the treatment group; and so on until all participants have been assigned. If the experiment has three conditions, then the search is for the numbers 1, 2, or 3 (or 1–3 for the first condition, 4–6 for the second, 7–9 for the third, ignoring 0); and the logic extends to any number of conditions. Such a procedure is called **simple random assignment** without replacement. It is "simple" because no prior stratification has occurred, and it is "without replacement" because any unit already assigned will be skipped if and when its number comes up a second time in the random number listing.

Restricted Random Assignment to Force Equal Sample Sizes

Simple random assignment often results in unequal sample sizes in each condition, especially when the sample of units is small. In the psychotherapy example, it is extremely unlikely that simple random assignment will split the 20 clients into two groups of 10 each. The reason is simple: in any random number list, it is extremely unlikely that one will encounter exactly 10 ones and 10 twos among the first 20 eligible numbers. More likely one will encounter 9 ones and 11 twos or 12 ones and 8 twos. With small samples, splits can be very uneven. One of us (Shadish) once accidentally demonstrated this to a class of graduate students by using the flip of a coin to publicly assign 10 students in the class to one of two conditions. The coin came up heads all 10 times, resulting in a treatment group with 10 participants and a control group with none! Unequal sample sizes can complicate the estimation of effects (Keppel, 1991), make the statistical analysis more sensitive to departures from homogeneity of variance (S. Maxwell & Delaney, 1990), and jeopardize power as the split exceeds a ratio of 2:1 or 3:1 (Pocock, 1983). Researchers should probably avoid simple random assignment with total sample sizes less than 200. Pocock (1983, Table 5.3) has shown with 200 units that a split as large 57–43% would be expected by chance only 5% of the time. The researcher must also consider whether interim results are desired. Unequal sample size splits are more likely early in a trial if strict procedures are not followed to ensure otherwise. If the researcher knows that interim results on less than the full sample will be computed or disseminated to the press, the funding agent, service providers, participants, or other scientists, then smaller group sizes are desirable to ensure equal sample sizes early on.

With smaller samples or when interim results are to be used, the researcher should consider methods that equalize sample sizes across conditions. One way to

do so is to randomize from within groups of size equal to the number of conditions—from pairs in studies with two conditions, from triplets in experiments with three conditions, and so on. With two conditions, the researcher takes the first two units, randomly assigns one to the first condition, and then places the second unit into the second condition. Often this method is used when the design calls for matching or stratifying on some predictor of outcome, a topic we cover later in this chapter. Whether or not matching is done on a substantive variable, masking of random assignment may then be required to prevent those involved in the experiment from using knowledge of assignment from matches to anticipate the next assignment. Thus in medical trials a physician who knew that matches of two were being used might be able to predict the next assignment, systematically directing patients toward or away from being the next treatment to be assigned, for example, by claiming that the next person to be assigned does not meet the eligibility criteria. The researcher can also keep the size of the matches confidential or even vary the size of the matches, though this would complicate analyses. Berger and Exner (1999) present a method for detecting selection biases that might have occurred as a result of such anticipation. However, when the experiment uses several stratification variables, has multiple factors with several levels each, or is a double-blind or multisite experiment, then the chances of being able to predict the next assignment are much smaller.

When units become available for treatment slowly over time—that is, when they "trickle" into a study—an adaptive assignment design is another option to equalize sample size over time (Kadane, 1996). Here the proportion assigned to conditions is changed over time to assign more units to the condition that has the smallest number of units. Like matching or stratifying, this method can also be used to equalize groups on substantive variables that have been measured prior to assignment, and this method also makes it harder for those involved in the experiment to predict the next assignment, so that the need to mask assignment is smaller.

The least desirable solution usually occurs when the researcher has not anticipated this problem at the start—to force equal cell sizes by starting with simple random assignment, to stop assigning units into a condition when it receives its planned share, and to continue in this way with all the other treatments until the last units are placed into the last remaining condition. The problem here is that this procedure increases the possibility of confounding if the last units differ systematically from earlier units (e.g., due to seasonality effects). In a trial of a program to prevent smoking, for example, the last people to volunteer may be systematically less motivated to quit than the first people who eagerly sought the opportunity. In a trial of a new cancer chemotherapy, eligibility criteria may be strict for early clients; but if the desired sample size seems unattainable in the time available, the researcher may loosen eligibility criteria to hasten inclusion into the study, with the result that the last patients may be different from the first ones. If these last participants are disproportionately assigned to the last condition so as to force equal sample size, significant bias may result.

Restricted Random Assignment to Force Unequal Sample Sizes

Randomization does not require equal sample sizes. In fact, a common definition of random assignment, that participants must have an equal chance of being in all conditions, is wrong. All random assignment requires is assignment by chance, with each eligible unit having a nonzero chance of assignment to each condition. Procedurally, forcing unequal sample sizes is easy. For example, to assign two thirds of the participants to one condition and one third to the other with a table of random numbers, one could use the numbers 1 through 6 to assign to the first condition and 7 through 9 for the second, without using 0. The options here are obvious enough to need little elaboration.

Unequal sample sizes are useful for two reasons. One is statistical. Although common belief is that equal sample sizes increase power, this is not always true. In optimal design theory (e.g., McClelland, 1997), unequal assignment of observations to conditions can increase power to detect certain effects. For example, when comparing more than two conditions, power for some contrasts may be maximized with a weighted sum of detectable differences, in which the weights reflect unequal sample sizes. To know if unequal sample sizes increase power, researchers must identify the basic design (number of factors and levels per factor) and know which effects are of most interest (e.g., linear versus higher order polynomial; main effects versus interactions; nonordinal effects) in more detail than most social scientists are accustomed to (Atkinson & Donev, 1992; Mead, 1988; Orr, 1999). McClelland (1997) suggests some practical guidelines. First, use the same number of levels of an independent variable as it requires parameters to estimate the effect: two levels for a linear effect, three for quadratic, and so on. Second, consider an extra level to guard against underfitting the model: for example, add a third level to guard against a quadratic effect when a linear effect is hypothesized. Third, allocate observations disproportionately to extreme levels of the variable if nonlinear effects are likely. Fourth, in those rare cases in which random sampling of levels is used, power is maximized in analysis by using as few polynomial and interaction terms as necessary or by using one-degree-of-freedom focused tests for categorical variables.

The second reason to use unequal sample sizes is to respond to practical demands. For example, when only a small number of participants can be given a treatment but the pool of applicants is large, using more control participants can improve statistical power compared with equal sample size (Kish, 1987). Similarly, when contrasting two treatments with each other and with a control group, primary interest sometimes lies in the contrast between the two treatments, for instance, if both treatments have often been compared with control in the past studies. In this case, a smaller control group may be warranted. Another example is to compare a well-known treatment with a new treatment when precise estimates of the effects of the latter are needed so that more participants may be assigned to the latter. In another case, one can maximize the number of people receiving treatment for ethical reasons, as when Klesges, Haddock, Lando, and Talcott (1999)

assigned 75% of new Air Force trainees to a program to help them maint smoking cessation and only 25% to the control. In yet another instance, treatm may be more expensive than control, and the optimum distribution of a fix budget over treatment and control is at issue (Orr, 1999; Zucker et al., 1995). Su pose the experimenter has a fixed $50,000 budget. It costs $1,000 to includ treatment participant who receives an expensive medical regimen and outco tests, but only $250 to include a control participant who receives just the outco tests. If the investigator opts for an equal sample size of 40 per cell, with an eff size of $d = .50$ and $\alpha = .05$, power is .60. But if the same resources are allocat to 30 treatment participants and 80 controls, power rises to .64. However, pow drops dramatically as sample sizes become very discrepant. With 25 treatm participants and 100 controls, power is back to .60; and as cell sizes become ev more discrepant, power drops rapidly. Power is more complex with designs to t multiple conditions or interactions, in which case consulting a statistical exper good advice.

Batch Randomization

Batch randomization occurs when only small groups of units are available for signment to conditions at any one time (Reicken et al., 1974), as with a new c cer trial in which the flow of eligible patients is slow and out of the experimente control. The experimenter cannot wait months or years to obtain a complete l of all patients before assigning them to conditions, because some of them wo die and the rest would undoubtedly find another physician. So assignment de sions might be made for each batch of patients periodically. The experimen must then decide ahead of time what proportion of participants are to be assign to each condition and arrange a randomization procedure that anticipates th proportions and applies them consistently to each batch. The previously descrit procedures can all be adapted to these conditions, but keeping track of batche essential for later analyses of differences over batches.

Trickle Process Randomization

Trickle process randomization occurs when units trickle in slowly and must be signed immediately for practical or ethical reasons (e.g., Goldman, 1977)—for ample, psychotherapy clients whose symptoms require immediate help. He fixed proportions of units are randomly assigned to conditions based on the pected proportion of treatment group openings for new clients. Thus, in an in tient medical study, the experimenter might know that about 10 eligible patie arrive each week and that about four beds are available each week. Then a fix proportion of about 40% of patients are randomly assigned to treatment ea week. Braught and Reichardt (1993) review trickle process assignment and scribe a computer protocol for implementing it.

Adaptive Randomization Methods

These methods adapt the assignment proportion over time to correct imbalances in sample size (or covariate levels) when the initial randomization is not producing the desired ratio of participants in each condition. This happened in the National Job Training Partnership Act (JTPA) study, in which staff found it difficult to recruit enough eligible youth to fill program slots. Halfway through the study they decreased the proportion of eligibles assigned to the control group from a 1:2 control:treatment ratio to a 1:6 ratio at some sites (Orr, 1999). The rationale and procedure were outlined in Efron (1971), adapted to batch or trickle process assignment in L. Freedman and White (1976) and Pocock and Simon (1975), and adapted to random assignment from strata (under the rubric of urn randomization or play-the-winner randomization) in Lachin (1988), Wei (1978), and Zelen (1974). Palmer and Rosenberger (1999; Rosenberger, 1999) review the strengths and weaknesses of all these methods.

In the simplest case, when an imbalance is noted, the probability of assignment of subsequent participants is adjusted until the originally envisioned proportions are achieved. Then the original assignment proportions are used again. For example, in a trial with a desired 50:50 proportion over two groups but in which an imbalance has developed, Pocock (1983) suggests shifting to a new ratio of 2:3 if the trial is small (or 3:2, depending on the direction of the imbalance to be remedied) or to 3:5 for larger trials of 100 units or more. The NIMH Collaborative Depression project used adaptive randomization to remedy a developing imbalance over conditions in the assignment of both males and minorities (Collins & Elkin, 1985). In a more complex case (often called urn randomization), the adjustment occurs after each assignment—after each participant is assigned to one group, the probability of assignment to that group is decreased and is increased to the other group(s). The result is a long-run balance of sample size.

However, some kinds of adaptive assignment can cause problems. If the first 50 participants were assigned using simple random assignment and resulted in a 15:35 split, and if the remaining 50 participants were assigned with a different proportion to remedy this split, then bias could result if the last 50 participants were systematically different from the first 50. For instance, they might have been subject to different eligibility standards or to different seasonality effects. During the 3-year San Diego Job Search and Work Experience Demonstration experiment, the San Diego job market tightened, so later applicants differed from earlier applicants in some job-related characteristics (Greenberg & Shroder, 1997). In such cases, pooling applicants assigned before the change in proportion with those assigned after it can create biases because the different kinds of applicants in the pools before and after the change are assigned to groups in different proportions. If the investigator recognizes from the start of the project that imbalances may arise, urn randomization would be far less subject to this problem. Otherwise, one could randomly discard cases separately from treatment and control conditions for each batch, so that each batch contributes equal proportions of

experimentals and controls; but this is only viable with sample sizes large enough to avoid too much loss of power. Finally, batches can be analyzed as strata in the statistical analysis.

Haphazard Assignment

Some investigators substitute haphazard assignment, that is, a procedure that is not formally random but has no obvious bias, for random assignment. For example, alternating assignment was used in an evaluation of a pretrial arbitration procedure in Connecticut courts (Lind, 1985). In the Illinois On-Line Cross-Match Demonstration experiment, applicants with odd last-digit identification numbers were experimentals, whereas those with even last-digit identification numbers were controls (Greenberg & Shroder, 1997). In the Teen Outreach program experiment, some sites assigned by choosing every other name on an alphabetized list (Allen, Philliber, Herrling, & Kuperminc, 1997). Some haphazard procedures approximate random assignment reasonably well (Mosteller, Gilbert, & McPeek, 1980; McAweeney & Klockars, 1998), and if random assignment is not feasible, they may be a good choice. Even when haphazard assignment procedures may have a bias, they can be a source of stronger quasi-experiments than letting participants choose their own conditions (Staines et al., 1999).

However, whenever haphazard assignment is feasible, some form of random assignment is usually feasible, too. It makes little sense to substitute the former for the latter, because the latter's characteristics are so widely studied and understood, whereas the former's characteristics are not. Haphazard procedures can *appear* to be random when they actually have selection biases. For example, if two clients come to a psychotherapy clinic for therapy each day, one in the morning and one in the afternoon, alternating assignment will be biased toward using "morning persons" in the treatment group and "afternoon persons" in the control group Such persons may differ systematically from each other in ways we rarely fully understand, in circadian rhythms or sleep patterns, for example. Similarly, some haphazard methods such as alternating assignment allow one to anticipate which treatment will be given next; a person knows that because the date is even-numbered, the patient will be assigned to control rather than treatment. If so, that person might encourage a particularly needy patient to delay coming until a day on which he or she can be assigned to treatment, thereby biasing the treatment group toward more needy patients. Because the possibility of such selection biases can never be fully ruled out in haphazard assignment, formal random assignment is usually preferred when it is feasible.

Sometimes investigators begin their studies with random assignment but later override randomization with a nonrandom procedure to compensate for an observed difference between groups. For example, the original Perry Preschool Project evaluation randomly assigned 128 children from 64 matched pairs to two conditions (Schweinhart et al., 1993). Subsequently, some of these pairs were exchanged so that groups would be better balanced for sex and socioeconomic sta-

tus (Spitz, 1993; Zigler & Weikart, 1993). This is tempting because it appears to do exactly what randomization intends—to make groups more equivalent before treatment begins. It is particularly tempting when the imbalance is on a variable known to predict outcome. However, it is not recommended. Differences between groups on observed means are to be expected because random assignment equates on expectations, not on observed means; and overriding randomization can even create biases on unobserved variables. If variables such as age are known to be important prognostic indicators, then the investigator should have done random assignment from matches or strata formed by age. It is rarely a good idea to override random assignment, for there are almost always better options, such as those we cover now.

What to Do If Pretest Means Differ

It is not a "failure of randomization" if some observed means are significantly different across conditions at pretest after random assignment.[7] Indeed, from sampling theory we expect that the fewer the number of participants assigned to conditions, the larger are the differences in pretest means that may be found. Conversely, large sample sizes greatly reduce the likelihood of substantial differences among observed pretest means and so enhance the interpretability of any single randomized trial. Similarly, using units that are very homogeneous also decreases the likelihood of such differences emerging. But as long as randomization was implemented properly, any pretest differences that occur will always be due to chance, even with small sample sizes, heterogeneous populations, and obtained pretest means that are greatly discrepant.

However, the fact that balance is expected in the long run over multiple experiments is little comfort in an experiment that is imbalanced here and now. One has to interpret the results of one's own experiment, imbalances and all. Hence researchers sometimes test whether pretest differences between conditions are statistically significant (or fall into a certain acceptable range; Hauk & Anderson, 1986; Makuch & Simon, 1978; Westlake, 1988). When these tests suggest that the observed pretest means differ by more than seems desirable, the researcher may then wonder whether to proceed with the experiment or to redo the randomization procedure until pretest means are observed that are more nearly equal. Unfortunately, redoing random assignment compromises posttest hypothesis tests whose underlying sampling distributions are based on the assumption of a single randomization. To rerandomize makes standard hypothesis testing a bit more conservative, failing to detect real effects too often. But some methodologists do allow rerandomization (e.g., Cox, 1958; Pocock, 1983) if the criteria under which

7. Huitema (1980) and Mohr (1995) call this unhappy randomization; Maxwell and Delaney (1990) call it fluke random assignment.

it will occur are specified in advance and if it is done prior to the start of the experiment. After all, even if rerandomization results in conservative tests, such conservative biases are often viewed as acceptable in science.

However, in most cases, there are better options. One can legitimately use the analysis of covariance (Huitema, 1980) to adjust results for random pretest differences, though this procedure is usually inferior to choosing covariates that are correlated with outcome (Begg, 1990; Maxwell, 1993; Permutt, 1990). Using large sample sizes also helps prevent pretest discrepancies of large magnitude from occurring, though the increased power they provide also increases the likelihood of finding significant differences on pretreatment measures. Even better, unhappy randomizations can be prevented from occurring at all through the use of random assignment from matches or strata, to which we now turn.

Matching and Stratifying

Units can be matched or stratified before being randomly assigned to conditions. For example, in an experiment on the effects of psychotherapy on stress, participants will differ from each other on initial stress levels and a host of other variables, such as marital status, that may be related to stress. Simple random assignment may yield observed mean differences between treatment and control groups on any of these variables. Such differences can be minimized by first matching participants on a relevant variable and then randomly assigning from these matches. For example, to match on age in a two-group experiment, the two oldest participants form a pair; and one member of the pair is randomly assigned to treatment and the other to control. Then the next oldest pair is treated similarly, and so on through the youngest pair. Diament and Colletti (1978), for example, used a behavioral counseling group and a wait-list control group to study the effects of a treatment of parental control problems with learning disabled children. Prior to the study, they placed 22 mothers into 11 pairs based on age and then randomly assigned from those pairs. Better still, the experimenter can match on pretest scores when they are available. Bergner (1974), for example, randomly assigned 20 couples to either marital therapy or a control group by using pairs formed by matching on pretest scores of behavior ratings of couple communication patterns, and similar behavior ratings at posttest were the outcome variable.

Stratifying is similar except that strata contain more units than conditions, unlike matches that contain the same number of units as conditions. For example, we might stratify psychotherapy clients on gender, randomly assigning females to treatment and control separately from males; or, in a multisite experiment, sites could be strata with random assignment to treatment and control within each site, as was done in the NIMH Treatment of Depression Collaborative Research Program (Elkin, Parloff, Hadley, & Autry, 1985; Imber et al., 1990). When either matching or stratifying can be used, matching is preferred if it is convenient, but stratifying into five strata is often sufficient to remove over 90% of the bias due

to the stratification variable (Cochran, 1968). Although we primarily discuss matching here, our comments also apply to stratifying unless otherwise indicated.

Both matching and stratifying greatly increase the likelihood that conditions will have similar pretest means and variances on the matching/stratifying variable and on any variables correlated with it. When properly analyzed, the variance due to the matching variable can then be removed from overall error variance (Keppel, 1991; Kirk, 1982). This produces a more powerful statistical test.[8] In general, we recommend random assignment from matches or strata whenever it is feasible and a good matching variable can be found. We especially recommend it when sample sizes are so small that simple random assignment may not result in equal means, variances, and sample sizes. This is important when subanalyses are planned (e.g., stratifying on gender to aid a test for a treatment by gender interaction). Without matching, there may be too few participants in individual cells to do the planned subanalyses.

We note three caveats. First, matching on variables completely unrelated to outcome produces little benefit and can decrease statistical power by using up degrees of freedom. Second, more care must be taken with matching in quasi-experiments, in which the issues we outlined in Chapters 4 and 5 come to bear (Campbell & Erlebacher, 1970; Costanza, 1995). Third, the benefits of matching and stratifying mostly apply to their use *prior to* randomization (Fleiss, 1986). For example, prestratification on gender with subsequent random assignment is far more likely to yield equal sample sizes than will post-randomization stratification on gender during data analysis.

A good matching variable in randomized experiments is often the pretest score on the main outcome of interest. So in the psychotherapy-for-stress example, matching participants on pretest stress levels prior to random assignment is preferred. If that is not possible, then the researcher can match on variables that are highly correlated with the outcome variable. For example, the researcher may not want to give a pretest on the exact same stress test to be used at posttest for fear of sensitizing clients to the outcome. But a simple one-item question about overall stress levels that is embedded in the initial application form that clients complete might be surreptitious enough to avoid sensitization effects but still be correlated with outcome. Or if it is known that overall stress levels are lower in married than in single clients, then stratifying on marital status may be worthwhile.

Adaptive randomization strategies can also be used to help equate groups on observed scores, in addition to their benefits for equating on sample size that we described earlier (e.g., Wei, 1978). The Project MATCH Research Group (1993)

8. There are two main exceptions. First is the case in which the number of units is very small (e.g., less than 10 per condition), as when schools or communities are matched and then assigned to conditions (Diehr, Martin, Koepsell, & Cheadle, 1995; Gail et al., 1992; Gail et al., 1996; Martin, Diehr, Perrin, & Koepsell, 1993). Second, if the between-match variance is not much larger than zero (e.g., if males do not differ appreciably from females on the outcome variable), then a randomized design without matching is more powerful because its error term will have more degrees of freedom.

used this method to ensure that the conditions in their alcohol treatment study were balanced for current drinking severity, prior treatment for drinking, prior psychiatric treatment, sociopathy, marital and employment status, gender, and education.

Finally, the Propensity Matched Pairs Design (PMPD) uses propensity score matching in randomized experiments (Hill, Rubin, & Thomas, 2000).[9] It can be used if a small proportion of a population is selected randomly to receive treatment, but there are insufficient resources to follow all remaining controls in the population. If sufficient pretreatment information is available to compute propensity scores on all treatment and nontreatment participants in the population, then a control can be selected from the population for each treatment participant by matching on propensity scores. This requires enumeration of the entire population of eligibles before the study starts. Presumably, methods for selecting multiple controls from the population for each treatment participant (e.g., Henry & McMillan, 1993; Rosenbaum, 1995a) could be adapted to this design.

Matching and Analysis of Covariance

The analysis of covariance (ANCOVA) can be used as an alternative or as a complement to matching or stratifying (Huitema, 1980; Maxwell & Delaney, 1990). ANCOVA is especially useful to adjust for variables that were not used for matching because their correlation with outcome was not anticipated or because it was too logistically complex to match on all variables that were known to predict outcome.[10] For example, in the NIMH Treatment of Depression Collaborative Research Program experiment (Elkin et al., 1985; Elkin et al., 1989; Imber et al., 1990), marital status was included as a covariate in the analysis because it proved to be correlated with outcome, though this was not anticipated. Similarly, Bloom (1990) used age, education, ethnicity, socioeconomic status (SES), earnings, employment, and unemployment benefits received as covariates in his study of the effects of reemployment services on the earnings and employment of over 2,000 displaced workers randomly assigned to conditions. The researcher could also covary propensity scores to adjust for random selection differences between conditions.

9. The true propensity for randomization into two conditions is a probability of .50 for each condition with simple random assignment. However, randomization equates conditions only on expectations, not on observed scores, so randomization does not equate observed propensity scores either. Predicting group membership in a randomized experiment will yield propensity scores that vary randomly from the true propensity scores of .50. Those propensity scores can then be used to adjust the randomized experiment for random assignment sampling fluctuations.

10. Researchers frequently make the mistake of choosing covariates that are significantly different between groups, but it is usually better to covary variables that are highly correlated with the outcome whether or not they distinguish between groups at pretest (Begg, 1990; Maxwell, 1993; Permutt, 1990). Beach and Meier (1989) present a method that takes into account the correlation of the covariate with both assignment and outcome.

The temptation is to add many covariates, but each uses up a degree of freedom. So adding covariates that do not predict outcome or that are highly correlated with each other will waste a degree of freedom while providing little new information.[11] Covariates also cost money to administer and score. Allison (1995) provides an algorithm for judging when the cost of adding a covariate exceeds the savings realized from needing fewer participants; Allison, Allison, Faith, Paultre, and Pi-Sunyer (1997) cover power-cost optimization with covariates, multiple measures, optimal allocation of participants to conditions, and optimal selection of participants to enter the experiment.

ANCOVA may be preferable to matching or stratifying on the covariate for control of extraneous variance due to the covariate if the relationship between covariate and outcome is linear or if the researcher is confident that the form of any nonlinearity can be adequately modeled. Matching and stratifying are not affected by such nonlinearities and so may be preferable if the relationship between the covariate and outcome is not known to be nonlinear or if the form of the nonlinearity is not confidently known (Maxwell & Delaney, 1990; Maxwell, Delaney, & Dill, 1984). Many statisticians are not confident that such nonlinearities can be well-modeled (Dehejia & Wahba, 1999). Matching may also be preferable if the matching variable has many small unordered categories, such as "type of alcoholic." Or matching and ANCOVA can be used together by randomly assigning from matches and then analyzing with ANCOVA; this allows all the benefits of matching for similar sample sizes and observed means at pretest.

The Human Side of Random Assignment

To judge from anecdotal evidence and research, problems in executing random assignment are prevalent (Berk, Smyth, & Sherman, 1988; Boruch & Wothke, 1985; Conrad, 1994; Dunford, 1990; Gilbert, McPeek, & Mosteller, 1977a; Greenberg & Shroder, 1997; Marcus, in press; Rezmovic, Cook, & Dobson, 1981; Sackett, 2000; Schulz, 1995; W. Silverman, 1977; Test & Burke, 1985).[12] Conner's (1977) analysis of 12 projects found that planned randomization was less likely when (1) randomization was done by researchers working for the project being evaluated rather than by outside researchers; (2) random assignment was controlled by operating personnel from the agency under study rather than by researchers; (3) loopholes exempted some individuals from random assignment; and (4) randomization was done by several persons rather than one individual. Similarly, Dennis (1988) found

11. The interpretation of ANCOVA results is complicated if the covariate is caused by treatment, a result that should generally be avoided by, for example, using a covariate measured prior to treatment beginning. If not, adjusting for the covariate may remove some of the treatment effect (Maxwell & Delaney, 1990, pp. 382–384).

12. The *Canadian Medical Association Journal* recently started a new series on why randomization fails (Sackett & Hoey, 2000).

that covert manipulation of randomization occurred often in 30 randomized experiments in criminal justice, especially when it was not controlled by the researcher or when the person doing the assignment knew to which conditions specific individuals were being assigned. These are the sorts of problems that field researchers often face. We know too little about their prevalence and about their likely influence on estimates of program effects, though available evidence suggests they can cause significant biases in experiments (Berger & Exner, 1999; Chalmers et al., 1983; Schulz, Chalmers, Hayes, & Altman, 1995).

Fortunately, we have learned much about how to increase the likelihood of successfully implementing random assignment (Boruch, 1997; Gueron, 1999; Orr, 1999). Boruch and Wothke (1985) suggest seven lessons that provide a convenient framework (Table 9.3). First, the researcher should plan how to explain the nature and purpose of randomization to those who will be affected, how to respond to arguments about why randomization could or should not be done, and how to provide incentives for doing randomization (Kruse et al., 2000). After all, it is naive to expect that research assistants, service delivery agents, program managers, or research participants will understand random assignment fully. The explanations need to be simple, scripted, learned in advance, and written and delivered at a level appropriate to the recipient. The researcher should anticipate the objections that are commonly raised to randomization and be prepared to discuss all sides of them, for such questions invariably arise and must be discussed hon-

TABLE 9.3 Seven Lessons About Implementing Random Assignment

1. Plan in advance how to explain the nature and purpose of randomization to those who will be affected, how to respond to various arguments about why randomization could or should not be done, and how to provide incentives for doing randomization.
2. Pilot test the randomization procedure to discover problems that can be remedied with further planning.
3. Develop clear procedures for implementing, controlling, and monitoring the randomization process throughout the entire experiment.
4. Have meetings at which to negotiate the randomization procedures with those who will be affected by them.
5. Develop fallback options that can be used to bolster estimates of program effects in the event that randomization fails.
6. Take advantage of naturally occurring opportunities that facilitate the conduct of randomization.
7. Carefully examine the match between the proposed design and those factors that will make randomization more likely to be successful in the particular context of the experiment.

Note: This table is a synthesis of a chapter on this topic by R. F. Boruch and W. Wothke, 1985.

estly and ethically. Gueron (1999) gives an example from an employment training experiment that required staff to randomly assign applicants to a control group:

> We met with these staff and told them what random assignment involved, why the results were uniquely reliable and believed, and how positive findings might convince the federal government to provide more money and opportunities for the disadvantaged youth they served, if not in San Jose, then elsewhere. They listened; they knew firsthand the climate of funding cuts; they asked for evidence that such studies had ever led to an increase in public funding; they sought details on how random assignment would work and what they could say to people in the control group. They agonized about the pain of turning away needy young people, and talked about whether this would be justified if other youth, as a result, gained new opportunities. Then they asked us to leave the room, talked more, and voted. Shortly thereafter, we were ushered back in and told that random assignment had won. This was one of the most humbling experiences I have confronted in 25 years of similar research projects, and it left me with a sense of awesome responsibility to deliver the study and get the findings out. (p. 11)

The worst mistake a researcher can make is not to take these staff and administrative concerns seriously. The ethical problems and the risk of failed randomization are too great. The experiment should often be abandoned if these concerns cannot be overcome.

Second, the researcher should pilot test the randomization procedure to discover problems that can be remedied. Gueron (1999) recommends that randomization schemes take no more than 1 minute per person to implement. Each person doing randomization should be instructed several times on how to randomize by an expert in randomization. In fact, there is probably no better way to find context-specific problems. For example, staff who have been cajoled into silence rather than persuaded of the value of random assignment may reveal their remaining objections through the actions they take as assignment is implemented. When piloting is not feasible, many implementation problems can be detected by using a brief "run-in" period during the start of the study in which procedures are tested but resulting data are not used in the analysis. If neither of these options is feasible, the researcher should pay particular attention to the initial randomization process to find and remedy problems before most experimental participants have been assigned.

Third, the researcher should develop clear procedures for implementing, controlling, and monitoring the randomization process throughout the entire experiment. It is usually better (1) to prepare the randomization procedures as early as possible, (2) to use tables of random numbers, except when public relations value requires mechanical methods, (3) to separate the randomization process from the process of determining eligibility so as to prevent covert manipulation of randomization by manipulation of eligibility judgments, (4) to have a single person who is part of the research team in charge of random assignment, (5) to keep the master list of assignments in a secure place (and keep a backup of it elsewhere), (6) to make that list accessible only to those doing assignment and to the principal investigator,

(7) to monitor the procedure closely and frequently for correct implementation throughout the experiment, (8) to check whether a new applicant has already been admitted into the study (because people often drop out and then reapply), (9) to keep a log of the assignment process and especially of violations of it, (10) to keep as many people as possible blind to assignment, (11) to have regular meetings with those doing randomization to review randomization decisions made since the last meeting and to identify and resolve problems with the procedure, and (12) where appropriate, to provide incentives to randomize correctly, such as monetary payments.[13]

Fourth, the researcher should have meetings at which to negotiate the randomization procedures with those who will be affected by them. Doing this when the randomization procedure is first designed allows the researcher to benefit from staff members' more intimate knowledge of what is likely to go wrong, to enlist their active cooperation in implementing randomization, and to minimize both the costs and the objections to randomization by those stakeholders. This includes those who will do randomization and also others in the organization who might be affected by randomization, such as service providers. Sherman and Berk (1985) note that implementing randomization is like implementing any other organizational change—such change always requires special attention. In the Minneapolis Spouse Abuse Experiment, Sherman and Berk spent considerable time in numerous meetings negotiating random assignment with the funding agency, the mayor of the city, representatives of the police who had to implement randomization, and various interest groups who cared about the problem. Through this negotiation procedure a feasible randomization procedure emerged. Some authors have even formalized these agreements in contracts (Bickman, 1985; Fairweather & Tornatsky, 1977).

Fifth, the researcher should develop fallback options that can be used to bolster estimates of program effects in the event that randomization fails. The addition of design elements such as staggered implementation of treatment over sites can provide alternative sources of causal evidence in the event that randomization fails in a serious way. Sixth, the researcher should take advantage of opportunities that facilitate the conduct of randomization. We covered a wide range of these opportunities in more detail at the end of the previous chapter. Seventh, the researcher should carefully examine the match between the proposed design and those factors that will make randomization more likely to be successful in the particular context of the experiment. For example, sometimes aggregate units such as classrooms or schools could be randomized more feasibly than individual units such as students, and the researcher should be prepared to adopt either design feature if possible. Conversely, the researcher should be skeptical of designs that rely on the promise that a site can generate a large number of experimental participants if that site has

13. Marcus (2001) discusses sensitivity analyses that can be helpful if it is suspected that randomization has been subverted.

not previously participated in experiments and can provide little evidence that large numbers of qualified participants exist and can be attracted.

All these matters are most difficult to ensure in multisite experiments in which participants within sites are assigned to conditions, particularly when a more complex design is used (Hausman & Wise, 1985). In such evaluations, the choice may be between allowing someone at the local level to randomize within each site or using a central office to randomize via telephone. The latter allows the researcher to retain control over the integrity of assignment but is more difficult to implement efficiently and quickly (Pocock, 1983). In multisite evaluations, different sites will also have different experience with randomization. Boruch et al. (1988) describe one site in a large multisite study that became distressed when all their referrals were assigned randomly to the control condition (stratifying by site prior to randomization can prevent this problem). Multisite evaluations require more complex randomization procedures that, in turn, require more resources to implement correctly, so researchers with few resources should minimize any complexity.

CONCLUSION

This chapter concerns what researchers can do to design an ethical experiment, to ensure that enough participants will be eligible to enter the experiment, and to plan a feasible and successful random assignment procedure. The next chapter takes up from this point and concerns problems that arise when those randomly assigned to treatment do not fully receive it or refuse to participate further in the measurement of their outcomes.

APPENDIX 9.1: RANDOM ASSIGNMENT BY COMPUTER

SPSS and SAS

The random number generators in programs such as SPSS and SAS are well-tested to produce reliable random assignments. Here is an SPSS syntax that will assign equal numbers of units randomly to conditions.[14] To run the program, begin by making the changes noted in the capitalized comments.

14. Both programs were suggested by Virgil Sheets of Indiana State University, used by permission.

```
input program.
*YOU SHOULD LOOP TO THE TOTAL N.
loop #I=1 to 200.
*DIVIDE THE N BETWEEN GROUPS.
if (#I<101) group=1.
if (#I>100) group=2.
compute x=normal(1).
end case.
end loop.
end file.
end input program.
sort cases by x.
print table/ $casenum group.
execute.
```

The number of cells in the design can be varied by adding more "if (#I>x) group=y" lines, and unequal cell sample sizes can be obtained by appropriate changes to the same statements. For example, to divide 200 units into three cells of 50, 50, and 100 each, use:

```
if (#I<101) group=1.
if (#I>100 and (#I<151)) group=2.
if (#I>150) group=3.
```

The following SAS syntax will also accomplish the task:

```
OPTIONS LS=80;
DATA WORK;
**Set number after 'to' as # in intended sample;
**change q to reflect # of groups in study;
**set intervals as equal division of sample by q;
DO I=1 TO 300;
IF I<61 THEN Q=1;
IF (I>60 AND I<121) THEN Q=2;
IF (I>120 AND I<181) THEN Q=3;
IF (I>180 AND I<241) THEN Q=4;
IF (I>240 AND I<301) THEN Q=5;
X=RANUNI(0);
OUTPUT;
END;
PROC SORT; BY X;
PROC PRINT; VAR Q;
```

World Wide Web

The World Wide Web contains numerous sites that can perform random assignment tasks. At the time of printing of this book, these sites included:
http://members.aol.com/johnp71/javastat.html#Specialized
http://www.assumption.edu/html/academic/users/avadum/applets/applets.html
http://lib.stat.cmu.edu/

Excel

Open a new workbook. Enter the number 1 in Cell A1 and 2 in A2; then drag the AutoFill handle (highlight Cells A1 and A2, and the handle appears as a dot at the lower right corner of A2) to Cell AN, where N is the number of units to be randomized. In B1, enter "=rand( )" (without the quotation marks); highlight B1 and drag the AutoFill handle to Cell BN. In Cell D1, enter the number 1, and drag the AutoFill handle to Cell Dx, where x is the number of units to be assigned to the first condition. In Cell D(x+1), enter the number 2, and drag the AutoFill handle to Cell D(x+y), where y is the number of units to be assigned to the second condition. If the study has a third condition, then in Cell D(x+y+1), enter the number 3, and drag the AutoFill handle to Cell D(x+y+z), where z is the number of units to be assigned to the third condition. Continue this process until all N cells in Column D are filled. Pick any cell in Column B and click the Sort Ascending button; this randomizes the original N units. Column A is the unit identification number, and Column D is the condition to which they are assigned. Highlight Columns B and C and delete them. Then either print the list (which is now ordered by condition number) or highlight Columns A and B and sort them in ascending order on Column A for a list ordered by unit identification number. Those who choose to use this procedure should note that errors in statistical procedures in Excel have been documented (McCullough & Wilson, 1999), at least for Excel 97; whether these errors have been fixed in subsequent versions is unknown. The nature of the errors is probably such that they would not compromise random assignment, but as other options do exist, those options should be used where feasible.