4

Research Design

Experiments and Experimental Thinking

IN THIS CHAPTER

INTRODUCTION

Early in the twentieth century, F. C. Bartlett, the pioneering psychologist who developed schema theory, went to Cambridge University to study with W. H. R. Rivers, an experimental psychologist. Rivers had been invited in 1899 to join the Torres Straits expedition and saw the opportunity to do comparative psychology studies of non-Western people (Tooker 1997:xiv). When Bartlett got to Cambridge, he asked Rivers for some advice. Bartlett expected a quick lecture on how to go out and stay out, about the rigors of fieldwork, and so on. Instead, Rivers told him: "The best training you can possibly have is a thorough drilling in the experimental methods of the psychological laboratory" (Bartlett 1937:416).

Bartlett found himself spending hours in the lab, "lifting weights, judging the brightness of lights, learning nonsense syllables, and engaging in a number of similarly abstract occupations" that seemed to be "particularly distant from the lives of normal human beings." In the end, though, Bartlett concluded that Rivers was right. Training in the experimental methods of psychology, said Bartlett, gives one "a sense of evidence, a realization of the difficulties of human observation, and a kind of scientific conscience which no other field of study can impart so well" (1937:417).

Whether you are doing questionnaire surveys, participant observation ethnography, or content analysis of texts, a solid grounding in the logic of the experimental method is one of the keys to good research skills. In this chapter, I discuss how experimental design and experimental thinking is used across the social sciences as a guide to better research on human thought and human behavior.

At the end of this chapter, you should understand the variety of research designs and how they are implemented in experiments, field research, and surveys. You should understand the concept of threats to validity and the various ways in which social scientists respond to those threats (**Further Reading:** research design).

EXPERIMENTS

There are several ways to categorize experiments. First of all, there is the distinction between randomized and nonrandomized assignment of participants, or true experiments versus quasi-experiments. In true experiments, participants (or subjects) are assigned randomly to either a treatment group or a control group. In quasi-experiments, subjects are selected rather than assigned.

Another way to categorize experiments is in terms of where they are done: in the laboratory or out in the world. Experiments in the lab offer greater control; field experiments offer greater realism. I distinguish two kinds of field experiments—natural experiments and naturalistic experiments—but the logic of experiments is the same no matter where they're done.

Randomized Experiments

Whether you're studying people or pigeons, doing research in the laboratory or in the wild, the rules for the design of any true experiment in the social sciences are the same as for experiments in physics or agriculture. There are, of course, differences in experiments with humans and experiments with objects or pigeons or plants. These differences, though, involve important ethical issues like deception, informed consent, and withholding of treatment, not logic. More on these ethical issues later.

THE LOGIC OF TRUE EXPERIMENTS

Steps in the Classic Experiment

There are five steps in a classic experiment:

1. Formulate a hypothesis.
2. Randomly assign participants to the intervention group or to the control group.

3. Measure the dependent variable(s) in one or both groups. This is called O_1 or "observation at time 1."
4. Introduce the treatment or intervention.
5. Measure the dependent variable(s) again. This is called O_2 or "observation at time 2."

Later, I'll walk you through some variations on this five-step formula, including one very important variation that does not involve Step 3 at all. But first, the basics.

Step 1.

Before you can do an experiment, you need a research question that can be studied using the experimental approach. In other words, you need a clear hypothesis about the relation between some independent variable (or variables) and some dependent variable (or variables). Experiments thus tend to be based on confirmatory rather than exploratory research questions (see Box 1.3).

The testing of new drugs can be a simple case of one independent and one dependent variable. The independent variable might be, say, "taking versus not taking" a drug. The dependent variable might be "getting better versus not getting better." The independent and dependent variables can be much more subtle. "Taking versus not taking" a drug might be "taking more of or less of" a drug and "getting better versus not getting better" might be "the level of improvement in high-density lipoprotein" (the so-called good cholesterol).

Move this logic to agriculture: *Ceteris paribus* (holding everything else—like amount of sunlight, amount of water, amount of weeding—constant), some corn plants get a new fertilizer and some don't. Then, the dependent variable might be the number of ears per corn stalk or the number of days it takes for the cobs to mature, or the number of grams of carbohydrates per cob.

Finally, move this same logic to human thought and human behavior: *Ceteris paribus*, police who take part in this new training program will be less aggressive in their arrests than will police who do not take part in it.

Things get more complicated when there are multiple independent (or dependent) variables. You might want to test two different training programs on police who come from three different ethnic backgrounds, for example. But the underlying logic for setting up experiments and for analyzing the results is similar across the sciences. When it comes to experiments, everything starts with a clear hypothesis.

Step 2.

You need at least two groups—the treatment group (also called the intervention group or the stimulus group) and the control group. One group gets the intervention (a new drug, for example, or exposure to a new method of teaching some subject), and the other group (the control group) doesn't. The treatment group (or groups) and the control group are involved in different experimental conditions.

In a true experiment, individuals are randomly assigned to either the intervention group or to the control group. This ensures that any differences between the groups are the consequence of chance and not of systematic bias. Some people in a population may be more religious, or more wealthy, or less sickly, or more prejudiced than others, but random assignment ensures that those traits are randomly distributed through the groups in an experiment.

Random assignment does not eliminate the possibility of selection bias altogether, but it makes differences between experimental conditions (groups) due solely to chance by taking the decision of who goes in what group out of your hands. The principle behind random assignment will become clearer after you work through Chapter 7 on probability sampling, but the bottom line is this: Whenever you *can*

assign participants randomly in an experiment, do it.

Step 3.

One or both groups are measured on one or more dependent variables. This is called the pretest.

Dependent variables in humans can be physical things like weight, height, number of leucocytes per milliliter of blood, or resistance to malaria. They can also be attitudes, psychological states, knowledge, or mental and physical achievements. For example, in weight-loss programs, you might measure the ratio of body fat to body mass as the dependent variable. If you are trying to raise women's understanding of the benefits of breast-feeding by exposing them to a multimedia presentation on this topic, then a preliminary test of women's attitudes about breast-feeding before they see the presentation is an appropriate pretest for your experiment. A preliminary score on a vocabulary test might be the pretest in an experiment to raise students' scores on the verbal part of the SAT.

Here are some social and psychological dependent variables I've seen recently in literature on experiments: attitude toward abortion; knowledge of mathematics among sixth graders; ability to function outside a hospital despite being clinically depressed; level of expressed racism; amount of self-esteem; level of perceived stress; support for the distribution of needles to intravenous drug users. In short, the dependent variable in an experiment can be anything you think might change as a result of some intervention.

You don't always need a pretest. More on this in a bit, when we discuss threats to validity in experiments.

Step 4.

The intervention (the independent variable) is introduced.

Step 5.

The dependent variables are measured again. This is the posttest.

A Walk-Through of an Example

Here's a made-up example of a true experiment: Take 100 college women (18–22 years of age) and randomly assign 50 of them to each of two groups. Bring each woman to the lab and show her a series of flash cards. Let each card contain a single, three-digit random number. Measure how many three-digit numbers each woman can remember. Repeat the task, but let the members of one group hear the most popular rock song of the week playing in the background as they take the test. Let the other group hear nothing. Measure how many three-digit numbers people can remember and whether rock music improves or worsens performance on the task.

Do you think this is a frivolous experiment? It turns out that many college students, ages 18–22, study while listening to rock music. It also turns out that this drives their parents crazy. I'll bet that more than one reader of this book has been asked something like: "How can you learn anything with all that noise?" The experiment outlined here is designed to test whether students can, in fact, "learn anything with all that noise."

There is plenty to criticize about this experimental design. Only women are involved and there are no graduate students, or high school students, either. Furthermore, there is no test of whether classic rock helps or hinders learning more than, say, hip-hop, or rhythm and blues, or country music, or Beethoven, or. . . . In fact, the experiment, as designed, doesn't even test whether people can learn anything important or *useful* when they listen or don't listen to rock music. The experiment tests only whether college-age women learn to memorize more or fewer three-digit numbers when the learning is accompanied by a single rock tune. The learning task is artificial.

But a lot of what's really powerful about the experimental method is embodied in this example. Suppose that the rock-music group does better on the task. We can be pretty sure

that it's not because of their gender or their age or their education, but because of the music. Just sticking in more independent variables (like expanding the group to include men, graduate students, or high school students; or playing different tunes; or making the learning task more realistic), without modifying the experiment's design to control for all those variables, creates what are called confounds to validity. They *confound* the experiment and make it impossible to tell if the intervention is what really caused any observed differences in the dependent variable.

Good experiments test narrowly defined questions. This is what gives them knowledge-making power. When you do a good experiment, you *know* something at the end of it. In this case, you know that women students at one school memorize or do not memorize three-digit numbers better when they listen to a particular rock tune.

I know this doesn't sound like much. Critics of experimental thinking in the social sciences say that it forces you to learn more and more about less and less, until finally you know everything there is to know about nothing.

Cute, but wrong. Think about it: The kind of knowledge you get from a well-designed experiment can be verified or falsified by another experiment. You can repeat the experiment at another school. If you get a different answer, then you need an explanation for this finding. Perhaps there is something about the student selection process at the two schools that produces the different results? Perhaps students at one school come primarily from working-class families, while students from the other school come from upper-middle-class families. Perhaps students from different socioeconomic classes grow up with different study habits, or prefer different kinds of music.

Conduct the experiment again, but include men this time. Conduct it again, and include two music conditions: a rock tune and a classical piece. Take the experiment on the road and run it all over again at different-sized schools in different regions of the country. Then, on to Paraguay . . .

INTERNAL AND EXTERNAL VALIDITY

True experiments, with randomized assignment and full control by the researcher, produce knowledge that has high internal validity. This means that changes in the dependent variables were probably *caused by*—not merely related to or correlated with—the treatment. Continual replication and verification produce cumulative knowledge, with high external validity—that is, knowledge that you can generalize to people who were not part of your experiment.

Replication of knowledge is every bit as important as its production in the first place. In fact, in terms of usefulness, replicated knowledge is exactly what we're after.

Consider the following experiment, designed to test whether offering people money produces fewer errors in an arithmetic task. Take two groups of individuals and ask them to solve 100 simple arithmetic problems. Tell one group that they will be given a dollar for every correct answer. Tell the other group nothing. Be sure to assign participants randomly to the groups to ensure equal distribution of skill in arithmetic. See if the treatment group (the one that gets the monetary rewards) does better than the control group.

This experiment can be embellished to eliminate confounds that threaten internal validity. Conduct the experiment a second time, with the same people, reversing the control and treatment groups. In other words, tell the former treatment group that they will not receive any financial reward for correct answers and tell the former control group that they will receive a dollar for every correct answer. (Of course, give them a new set of problems to solve.)

This creates a whole new experiment. You're no longer testing whether financial incentives motivate people to try harder in solving a set of arithmetic problems. Now you're also testing whether pulling financial incentives away from people who are accustomed to getting them will affect their ability to solve those arithmetic problems.

Conduct the experiment many times, changing or adding independent variables. In one version of the experiment, you might keep the groups from knowing about each other. In another, you might let each group know about the other's efforts and rewards (or lack of rewards). Perhaps when people know that others are being rewarded for good behavior and that they themselves are not rewarded they will double their efforts to gain the rewards (this is called the "John Henry effect"). Perhaps they just become demoralized and give up.

By controlling the interventions and the group membership, you can build up a series of conclusions regarding cause and effect between various independent and dependent variables.

While controlled experiments like these have the virtue of high internal validity, they have the liability of low external validity. It may be true that a reward of a dollar per correct answer results in significantly more correct answers for the groups you tested in your laboratory. But you can't tell whether a dollar is sufficient reward for all groups, or whether a quarter would be enough to create the same experimental results in some groups. Worst of all, you don't know whether the laboratory results explain *anything* you want to know about in the real world.

To test external validity, you might propose some kind of monetary reward for teaching children in a dozen actual third-grade classrooms, to do arithmetic—six classrooms in which, say, children earn a penny per correct answer, and six in which there is no monetary incentive. If it turns out that the kids in the intervention classrooms get higher scores on arithmetic tests at the end of the year, then the next question is: How far do the results generalize? Just to the classrooms in the experiment? To all third graders in the school district? To all third graders in the state? In the country?

Kinds of Confounds: Threats to Validity

It's pointless to ask questions about external validity until you establish internal validity. In a series of influential publications, Donald Campbell and his colleagues identified the threats to internal validity of experiments (see Campbell 1957, 1979; Campbell and Stanley 1966; T. D. Cook and Campbell 1979). Here are seven of the most important confounds:

History

The history confound refers to any independent variable, other than the treatment, that (1) occurs between the pretest and the posttest in an experiment and (2) affects the experimental groups differently. Suppose you are doing a laboratory experiment, with two groups (experimental and control), and there is a power failure in the building. So long as the lights go out for both groups, there is no problem. But if the lights go out for one group and not the other, it's difficult to tell whether it was the treatment or the power failure that causes changes in the dependent variable.

In a laboratory experiment, history is controlled by isolating subjects as much as possible from outside influences. When we do experiments outside the laboratory, it is almost impossible to keep new independent variables from creeping in and confounding things.

Recall the example of testing whether monetary incentives help third graders do better in arithmetic. Suppose that right in the middle of the school term during which the experiment was being conducted, the Governor's Task Force on Elementary Education issues its long-awaited report and it contains the observation

that arithmetic skills must be emphasized during the early school years. Furthermore, it says, teachers whose classes make exceptional progress in this area should be rewarded with 10% salary bonuses.

The governor accepts the recommendation and announces a request for a special legislative appropriation. Elementary teachers all over the state start paying extra attention to arithmetic skills. Even supposing that the students in the treatment classes do better than those in the control classes, how can we be certain that the magnitude of the difference would not have been greater had this historical confound not occurred?

Maturation

The maturation confound refers to the fact that people in any experiment grow older, or get more experienced while you are trying to conduct an experiment. Consider the following experiment: Start with a group of teenagers on a Native American reservation and follow them for the next 60 years. Some of them will move to cities, some will go to small towns, and some will stay on the reservation. Periodically, test them on a variety of dependent variables (their political opinions, their wealth, their health, their family size, and so on). See how the various experimental treatments (city vs. reservation vs. town living) affect these variables.

Here is where the maturation confound enters the picture. The people you are studying get older. Older people in many societies become more politically conservative. They are usually wealthier than younger people. Eventually, they come to be more illness-prone than younger people. Some of the changes you measure in your dependent variables will be the result of the various treatments and some of them may just be the result of maturation.

Maturation is sometimes taken too literally. Social service delivery programs "mature" by working out bugs in their administration. People "mature" through practice with experimental conditions and they become fatigued. We see this all the time in new social programs where people start out being enthusiastic about innovations in organizations and eventually get bored or disenchanted.

Testing and Instrumentation

The testing confound occurs in laboratory and field experiments when subjects get used to being tested for indicators on dependent variables. This quite naturally changes their responses. Asking people the same questions again and again in a longitudinal study, or even in an ethnographic study done over six months or more, can have this effect.

The instrumentation confound results from changing measurement instruments. Changing the wording of questions in a survey is essentially changing instruments. Which responses do you trust: the ones to the earlier wording or the ones to the later wording? If you do a set of observations in the field—like children's behavior at recess or nurses' behavior in responding to patients in a hospital or cops' behavior in making arrests—and later send in someone else to continue the observations, you have changed instruments.

Which observations do you trust as closer to the truth: yours or those of the substitute instrument (the new field researcher)? In multi-researcher projects, this problem is usually dealt with by training all investigators to see and record things in more or less the same way. This is called increasing interrater reliability. (More on this in Chapter 19, on analyzing qualitative data.)

Regression to the Mean

Regression to the mean is a confound that can occur when you study groups that have extreme scores on a dependent variable. No matter what the treatment is, over time you'd expect the extreme scores to become more moderate, just because there's nowhere else for them to go. If men who are taller than 6′7″

marry women who are taller than 6′3″, then their children are likely to be (1) taller than average and (2) closer to average height than either of their parents are. There are two independent variables (the height of each of the parents) and one dependent variable (the height of the children). We expect the dependent variable to "regress toward the mean," since it really can't get more extreme than the height of the parents.

I put that phrase "regress toward the mean" in quotes because it's easy to misinterpret this phenomenon—to think that the "regressing" toward the mean of an dependent variable is caused by the extreme scores on the independent variables. It isn't, and here's how you can tell that it isn't: Very, very tall children are likely to have parents whose height is more like the mean. One thing we know for sure is that the height of children doesn't cause the height of their parents. Regression to the mean is a statistical phenomenon—it happens in the aggregate and is not something that happens to individuals (Box 4.1).

Box 4.1 The problem of extreme values

Many social intervention programs make the mistake of using people with extreme values on dependent variables as subjects. Suppose you're asked to evaluate a new reading program for third graders. To give the program a real workout, you choose children who scored in the bottom 10% of their class and compare them to children in the top 10% of their class. You'll probably find that the bottom 10% shows improvement, but don't write home about this just yet. You'll also probably find that the top 10% shows some decrease in their performance.

What's going on? Well, those bottom 10-percenters have nowhere to go but up, and the top 10-percenters have nowhere to go but down. If you introduced no program at all, the kids at the extremes would have a statistical chance, just by random fluctuation, of getting scores that are more like the mean the next time you test them. By choosing groups that score at the extreme, you wind up not being able to tell if the new reading program caused the change in scores, or if the change was just a statistical regression to the mean.

Selection of Participants

Selection bias in choosing subjects is a major confound to validity in both quasi-experiments and natural experiments. In laboratory experiments, you assign subjects at random, from a single population, to both treatment groups and control groups. This distributes any differences among individuals in the population throughout the groups, making the groups equivalent. This reduces the possibility that differences among the groups will cause differences in outcomes on the dependent variables, so selection is not a threat to the internal validity of the experiment.

Random assignment of participants to experimental conditions *reduces* the possibility of selection bias, but it doesn't eliminate the possibility altogether. Random assignment, then, maximizes the chance for valid outcomes—outcomes that are not clobbered by hidden factors.

In natural experiments, we have *no control* over assignment of individuals to groups.

Question: Do victims of violent crime have less stable marriages than persons who have not been victims? Obviously, researchers cannot randomly assign subjects to the treatment (violent crime). It could turn out that people who are victims of violent crime are more likely to have unstable marriages anyway, even if they never experienced violence.

Question: Do migrants to cities from small towns engage in more entrepreneurial activities than stay-at-homes? If we could assign

rural people randomly to the treatment group (those engaging in urban migration), we'd have a better chance of finding out. Since we cannot, selection is a threat to the internal validity of the experiment. Suppose that the answer to the question at the top of this paragraph were "yes." We still don't know the direction of the causal arrow: Does the treatment (migration) cause the outcome (greater entrepreneurial activity)? Or does having an entrepreneurial personality cause migration?

Mortality

The mortality confound refers to the fact that individuals may not complete their participation in an experiment. Suppose we follow two sets of married couples for five years. The couples in one group get free family counseling sessions once every three months. The other couples don't. During the first year of the experiment we have 200 couples in each group. By the fifth year, 30 couples have dropped out of the treatment group and 130 of the 170 remaining couples (76%) are still married. In the control group, 50 couples have dropped out and 75 of the 150 remaining couples (50%) are still married. One conclusion is that lack of counseling caused those in the control group to get divorced at a faster rate than those in the treatment group.

But what of those 30 couples in the treatment group and the 50 couples in the control group who left? It could be that they were mostly still married or mostly divorced. In either case, this would affect the results of the experiment, but we just don't know. Mortality can be a serious problem in natural experiments if it gets to be a large fraction of the group(s) under study.

Mortality also affects panel surveys. That's where you interview the same people more than once to track something about their lives. (More about panel studies in Chapter 9.)

Diffusion of Treatments

The diffusion of treatments threat to validity occurs when a control group cannot be prevented from receiving the treatment in an experiment. This is particularly likely in quasi-experiments where the independent variable is an information program.

In a project with which I was associated some years ago, a group of African Americans were given instruction on modifying their diet and exercise behavior to lower their blood pressure. Another group was randomly assigned from the population to act as controls—that is, they did not receive instruction. The evaluation team measured blood pressure in the treatment group and in the control group before the program was implemented. But when they went back after the program was completed, they found that control group members had also been changing their behavior. They had learned of the new diet and exercises from the members of the treatment group.

CONTROLLING FOR THREATS TO VALIDITY

In what follows, I want to show you how the power of experimental logic is applied to real research problems. The major experimental designs are shown in Figure 4.1. The notation is pretty standard.

> **X** stands for some intervention—a stimulus or a treatment of a subject.
>
> **R** means that subjects are randomly assigned to experimental conditions—either to the intervention group that gets the treatment, or to the control group that doesn't.

Several designs include random assignment and several don't.

> **O** stands for "observation." $\mathbf{O}_1$ means that some observation is made at Time 1. $\mathbf{O}_2$ means that some observation is made at Time 2, and so on.

Observation means "measurement of some dependent variable," but as you already know,

Figure 4.1 Some Research Designs

Figure 4.1a The Classic Design: Two-Group Pretest-Posttest

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
Group 1	*R*	O_1	*X*	O_2
Group 2	*R*	O_3		O_4

Figure 4.1b The Solomon Four-Group Design

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
Group 1	*R*	O_1	*X*	O_2
Group 2	*R*	O_3		O_4
Group 3	*R*		*X*	O_5
Group 4	*R*			O_6

Figure 4.1c The Classic Design Without Randomization

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
Group 1		O_1	*X*	O_2
Group 2		O_3		O_4

Figure 4.1d The Campbell and Stanley Posttest-Only Design

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
Group 1	*R*		*X*	O_1
Group 2	*R*			O_2

Figure 4.1e The One-Shot Case Study Design

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
			X	*O*

Figure4.1f The One-Group Pretest-Posttest Design

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
		O_1	*X*	O_2

Figure 4.1g Two-Group Posttest-Only Design: Static Group Comparison

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
			X	O_1
				O_2

Figure 4.1h The Interrupted Time Series Design

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
		OOO	*X*	*OOO*

the idea of measurement is pretty broad. It can be taking someone's temperature or testing their reading skill. It can also be just writing down whether they are smiling.

The Classic Design

We begin with the classic experimental design, the two-group pretest-posttest with random assignment. It is shown in Figure 4.1a. From a population of potential participants, some participants have been assigned randomly to a treatment group and a control group. Read across the top row of the table. An observation (measurement) of some dependent variable or variables is made at time 1 on the members of group 1. That is O_1. Then an intervention is made (the group is exposed to some treatment, X). Then, another observation is made at time 2. That is O_2.

Figure 4.1a The Classic Design: Two-Group Pretest-Posttest

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
Group 1	*R*	O_1	*X*	O_2
Group 2	*R*	O_3		O_4

Now look at the second row of the Figure 4.1a. A second group of people are observed, also at time 1. Measurements are made of the same dependent variable(s) that were made for the first group. The observation is labeled O_3. There is no X on this row, which means that no intervention is made on this group of people. They remain unexposed to the treatment or intervention in the experiment. Later, at time 2, after the first group has been exposed to the intervention, the second group is observed again. That's O_4.

Random assignment of participants ensures equivalent groups, and the second group, without the intervention, ensures that several threats to internal validity are taken care of. Most importantly, you can tell how often (how many times out of a hundred, for example) any differences between the pretest and posttest scores for the first group might have occurred anyway, even if the intervention hadn't taken place.

The classic experimental design is used widely across the social sciences to evaluate education programs—everything from teaching dental students in how to handle a mirror in a patient's mouth (it's not easy; everything you see is backward), to teaching Samoan women the value of getting a Pap smear (see Kunovich and Rashid [1992] and Mishra et al. [2009]).

Patricia Chapman and colleagues (Chapman et al. 1997) wanted to educate student athletes about sports nutrition. They had access to an eight-team girl's high school softball league in southern California. The participants were 14–18 years old on each team. Chapman et al. assigned each of 72 players randomly to one of two groups. The girls in the treatment group got two 45-minute lectures a week for six weeks about things like dehydration, weight loss, vitamin and mineral supplements, energy sources, and so on. The control group got no instruction.

Before the six-week program started, the researchers asked each participant to complete the Nutrition Knowledge and Attitude Questionnaire (Werblow et al. 1978) and to list the foods they'd consumed in the previous 24 hours. The nutrition knowledge-attitude test and the 24-hour dietary recall test were the pretests in this experiment. Then, when the six-week program was over, Chapman et al. gave the participants the same two tests. These were the posttests. The pretests provided

baseline data and the posttests provided data for assessing whether the nutrition education program had made a difference.

The education intervention did make a difference—in knowledge, but not in reported behavior. The participants in both the treatment and control groups scored about the same on the knowledge/attitude test in the pretest. After they went through the lecture series, the participants in the treatment group scored about 18 points more (out of 200 possible points) than those in the control group. Before the program, the 36 girls in the control group reported an average 24-hour intake of 1,683 calories and those in the treatment group reported an average of 2,054 calories.

As Chapman et al. point out, though, even 2,054 calories is not enough for adolescent females who are involved in competitive sports. After the intervention—after all those lectures—the reported 24-hour average caloric intake was 1,793 for the control group and 1,892 for the treatment group. In other words, if the intervention had any effect on behavior, it was to lower (and hence, worsen) the intake of these young female athletes. Chapman et al. point out that the results confirm the findings of other studies: For many adolescent females, the attraction of competitive sports is the possibility of losing weight.

The Solomon Four-Group Design

The classic design has one important flaw: It is subject to testing bias. Differences between variable measurements at time 1 and time 2 might be the result of the intervention, but they also might be the result of people getting savvy about being watched and measured. Pretesting can, after all, sensitize people to the purpose of an experiment, and this, in turn, can change people's behavior. The Solomon four-group design, shown in Figure 4.1b, controls for this. Since there are no measurements at time 1 for groups 3 and 4, this problem is controlled for. (This design is named for Richard Solomon. See Solomon [1949] and Solomon and Lessac [1968].)

Larry Leith (1988) used the Solomon four-group design to study a phenomenon known to all sports fans as the "choke." That's when an athlete plays well during practice and then loses it during the real game, or plays well all game long and folds in the clutch when it really counts. It's not pretty.

Leith assigned 20 male students randomly to each of the four conditions in the Solomon four-group design. The pretest and the posttest were the same: Each participant shot 25 free throws on a basketball court. The dependent variable was the number of successful free throws out of 25 shots in the posttest. The independent variable—the treatment—was giving or not giving the following little pep talk to each participant just before he made those 25 free throws for the posttest:

> Research has shown that some people have a tendency to choke at the free-throw line when shooting free throws. No one knows why some

Figure 4.1b The Solomon Four-Group Design

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
Group 1	*R*	O_1	*X*	O_2
Group 2	*R*	O_3		O_4
Group 3	*R*		*X*	O_5
Group 4	*R*			O_6

people tend to choking behavior. However, don't let that bother you. Go ahead and shoot your free throws. [Leith 1988:61]

What a wonderfully simple, utterly diabolic experiment. You can guess the result: There was a significantly greater probability of choking if you were among the groups that got that little pep talk, irrespective of whether they'd been given the warm-up pretest.

The Solomon four-group design is very useful in evaluation research. From New York to Nairobi, social workers, teachers, and researchers have been looking for ways to teach adolescents about the effective use of condoms in the prevention of sexually transmitted diseases and unwanted pregnancies. Interventions are usually some kind of teaching program administered in the classroom, and study after study shows that they don't work.

Kvalem et al. (1996) evaluated one of these in-school programs. They had 124 classes of 16–20 year olds, comprising a total of 2,411 students from Vestvold County in Norway. They assigned the classes randomly to the four conditions in the four-group design. I mention this study because the major finding was not that the program worked. It didn't. The major finding was a strong interaction effect between the pretest and the intervention on the reported use of condoms.

The pretest was an 80-item questionnaire about sexual behavior, use of condoms, and demographics. The posttest was the same questionnaire, sent six months and 12 months after the intervention. Of the four groups on the Solomon four-group design, the group that had *both* the pretest *and* the intervention had a higher likelihood of reporting condom use six months after the intervention. Six months later, though, when the posttest was given again, the interaction effect had disappeared.

The study by Kvalem et al. is instructive. It shows clearly the importance of testing for the effects of pretesting in social and psychological experiments. And it shows clearly the importance of following up with a second posttest to make sure that any effects on the dependent variable have lasted. If Kvalem et al. had not done that second posttest, they might have been tempted to interpret the interaction effect as having policy implications: If you want to get adolescents to use condoms, combine a pretest with an educational intervention. Kvalem et al.'s second posttest questionnaire, sent to the participants after 12 months, stopped them from making that mistake.

The Two-Group Pretest-Posttest Without Random Assignment

Figure 4.1c shows the design for a two-group pretest-posttest without random assignment. In this quasi-experiment, participants are not assigned randomly to the control and the experimental condition. This compromise with design purity is often the best we can do.

Watkins et al. (2010) studied the healing power of the Vietnam Veterans Memorial (VVM). They had a population of 62 Vietnam vets who were in treatment at a VA hospital for post-traumatic stress syndrome (PTSD). Of those, 32 went on an annual excursion to the

Figure 4.1c The Classic Design Without Randomization

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
Group 1		O_1	X	O_2
Group 2		O_3		O_4

VVM. All the vets had their PTSD symptoms tested regularly. A month after the trip to the VVM, those who went on the trip, and who had been on the trip at least twice before, has less severe symptoms.

Campbell and Boruch (1975) showed that lack of random assignment in quasi-experiments like this can lead to problems. Suppose you invent a technique for improving reading comprehension among third graders. You select two third-grade classes in a school district. One of them gets the intervention and the other doesn't. Students are measured before and after the intervention to see whether their reading scores improve.

Suppose the children in one class are from wealthier homes, on average, than are the children in the other class and suppose that the wealthier children test higher in reading comprehension at the end of the year than the poorer children. Would you (or the agency you're working for) be willing to bet, say, $300,000 on implementing the new reading comprehension program in all classes in the school district? Would you bet that it was the new program and not some confound, like socioeconomic class, that caused the differences in test scores?

If all the children, one at a time, were assigned randomly to the two groups (those who got the program and those who didn't), then this confound would disappear—not because socioeconomic status stops being a factor in how well children learn to read, but because children from poor and rich families would be equally likely to be in the treatment group or in the control group. Any bias that socioeconomic status causes in interpreting the results of the experiment would be distributed randomly and would, in theory, wash out.

But children come packaged in classrooms. It's physically impossible for a teacher to administer two separate programs in one classroom, so evaluation of these kinds of interventions are usually quasi-experiments because they have to be.

The Posttest-Only Design With Random Assignment

Look carefully at Figure 4.1d. It is the second half of the Solomon four-group design and is a posttest-only design with random assignment. This design—known also as the Campbell and Stanley posttest-only design—has a lot going for it. It retains the random assignment of participants in the classical design and in the Solomon four-group design, but it *eliminates pretesting*—and the possibility of a confound from pretest sensitization. When subjects are assigned randomly to experimental conditions (control or treatment group), a significant difference on O_1 and O_2 in the posttest-only design means that we can have a lot of confidence that the intervention, X, caused that difference (T. D. Cook and Campbell 1979).

Another advantage is the huge saving in time and money. There are no pretests in this design and there are only two posttests instead of the four in the Solomon four-group design.

McDonald and Bridge (1991) used this elegant design in their study of how gender stereotyping by nurses affects how nurses care for patients.

Figure 4.1d The Campbell and Stanley Posttest-Only Design

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
Group 1	*R*		*X*	O_1
Group 2	*R*			O_2

McDonald and Bridge asked 160 female medical-surgical nurses to read an information packet about a colostomy patient whom they would be attending within the next 8 hours. The nurses were assigned randomly to one of eight experimental conditions: (1) The patient was named Mary B. or Robert B. to produce *two patient-gender conditions*. (2) Half the nurses read just a synopsis of the condition of Mary B. or Robert B. and half read the same synopsis as the fourth one in a series of seven. This produced *two memory-load conditions*. (3) Finally, half the nurses read that the temperature of Mary B. or Robert B. had just spiked unexpectedly to 102°, and half did not. This produced *two patient stability conditions*.

The three binary conditions combined to form eight experimental conditions in a factorial design (more on factorial designs at the end of this chapter).

Next, McDonald and Bridge asked nurses to estimate, to the nearest minute, how much time they would plan for each of several important nursing actions. Irrespective of the memory load, nurses planned significantly more time for giving the patient analgesics, for helping the patient to walk around, and for giving the patient emotional support when the patient was a man (Box 4.2).

Box 4.2 Posttest-only: The underappreciated design

The Campbell and Stanley posttest-only design, with random assignment, is named for Donald Campbell and Julian Stanley (1963). It is not used as much as I think it should be, despite its elegance, its delightful simplicity, and its low cost. In a recent search of the literature, I found 1,110 examples of studies that used the pretest-posttest design, compared to 133 for studies that used the posttest-only design (with or without random assignment). This preference for the classic design is due partly to the appealing-but-mistaken idea that matching participants in experiments on key independent variables (age, ethnicity, etc.) is somehow better than randomly assigning participant to groups and partly to the nagging suspicion that pretests are essential to the experimental method.

That nagging suspicion—that we can do better than trust the outcome of events to randomness—has been the focus of a lot of research since a paper by Gilovich et al. in 1985 titled: "The Hot Hand in Basketball—On the Misperception of Random Sequences." The hot-hand phenomenon—the belief that streaks (in sports and in money management, for example) are the result of nonrandom forces—is hard to break. By the same token, so is the belief that small samples, if drawn randomly, are sufficient to warrant generalizing to a population. On this one, see the 600+ citations to Tversky and Kahneman (1971) and Chapters 6 and 7 on representative and nonrepresentative sampling

The One-Shot Case Study

The one-shot case study design is shown in Figure 4.1e. It is also called the ex post facto design because a single group of individuals is measured on some dependent variable *after* an intervention has taken place.

This is the most common design in culture change studies, where it is obviously impossible to manipulate the dependent variable. You arrive in a community and notice that something important has taken place. A new exit on the highway produces more tourist traffic or an interstate highway has bypassed a town and tourism revenues have plummeted. You try to evaluate the natural experiment by interviewing people (O) and by trying to assess the impact of the intervention (X).

Figure 4.1e The One-Shot Case Study Design

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
			X	O

With neither a pretest nor a control group, you can't be sure that what you observe is the result of some particular intervention. Despite this apparent weakness, however, the intuitive appeal of findings produced by one-shot case studies can be formidable.

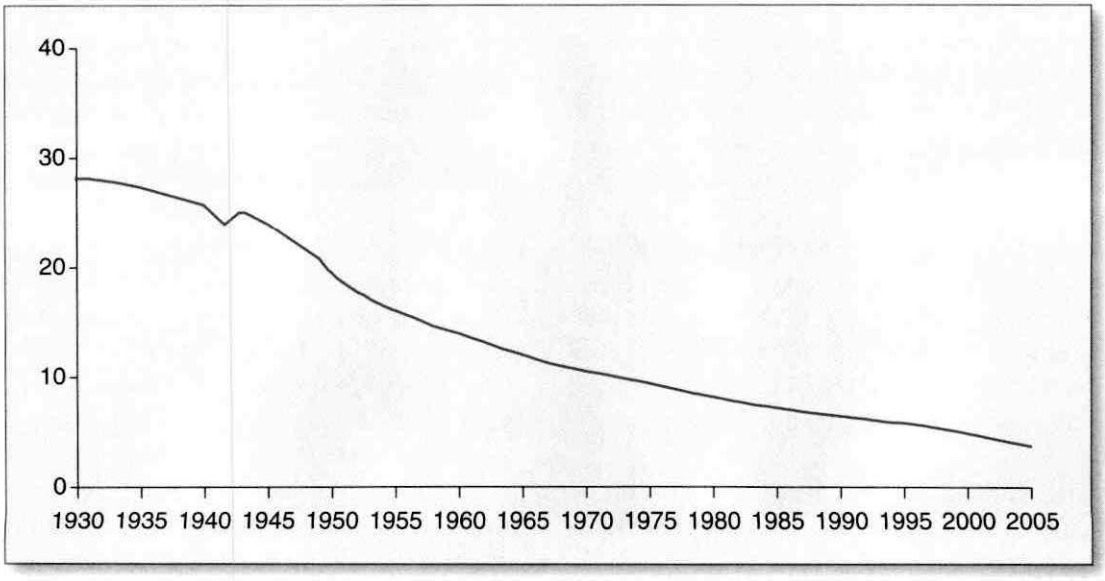
In the 1950s, physicians began general use of the Pap Test, a simple office procedure for determining the presence of cervical cancer. Figure 4.2 shows that since 1950, the death rate from cervical cancer in the United States has dropped steadily, from about 18 per 100,000 women to about 11 in 1970, to about 8.3 in 1980, to about 6.5 in 1995, and to about 2.4 in 2005. On the other hand, if you look only at the data *after* the intervention (the one-shot case study X O design) you might conclude that the intervention (the Pap Test) caused this drop in cervical cancer deaths. There is no doubt that the continued decline of cervical cancer deaths is due largely to the early detection provided by the Pap Test, but by 1950, the death rate had already declined by 36% from 28 per 100,000 women in 1930 (Williams 1978:16).

Never use a design of less logical power when one of greater power is feasible. If pretest data are available, use them. On the other hand, a one-shot case study is often the best you can do. Virtually all ethnography falls in this category, and as I have said before, nothing beats a good story, well told (**Further Reading:** case study methods).

The One-Group Pretest-Posttest

The one-group pretest-posttest design is shown in Figure 4.1f. Some variables are measured (observed), then the intervention takes place,

Figure 4.2 Death Rate From Cervical Cancer, 1930–2005



Source: Adapted from B. Williams, *A Sampler on Sampling*, Figure 2.1, p. 17. ©1978, Lucent Technologies.

Figure 4.1f The One-Group Pretest-Posttest Design

Time 1			Time 2	
	Assignment	Pretest	Intervention	Posttest
		O_1	X	O_2

Figure 4.1g Two-Group Posttest-Only Design: Static Group Comparison

Time 1			Time 2	
	Assignment	Pretest	Intervention	Posttest
			X	O_1
				O_2

and then the variables are measured again. This takes care of some of the problems associated with the one-shot case study, but it doesn't eliminate the threats of history, testing, maturation, selection, and mortality. Most importantly, if there is a significant difference in the pretest and posttest measurements, we can't tell if the intervention made that difference happen.

M. Peterson and Johnstone (1995) studied the effects on 43 women inmates of a U.S. federal prison of an education program about drug abuse. The participants all had a history of drug abuse, so there is no random assignment here. Peterson and Johnstone measured the participants' health status and perceived well-being before the program began and after the program had been running for nine months. They found that physical fitness measures were improved for the participants as were self-esteem, health awareness, and health-promoting attitudes.

The one-group pretest-posttest design is commonly used in evaluating training programs. The question asked is: Did the people who were exposed to this skill-building program (police officers, nurses, kindergarten teachers, high school algebra students, etc.) get any benefit out of it, and if so, how much?

The Two-Group Posttest Only Design Without Random Assignment

The two-group posttest only design without random assignment design is shown in Figure 4.1g. This design, also known as the static group comparison, improves on the one-shot *ex post facto* design by adding an untreated control group—an independent case that is evaluated only at time 2. In this design, however, the researcher has no control over assignment of participants to the intervention or control group. This creates an unresolvable validity threat.

There is no way to tell whether the two groups were comparable at time 1, before the intervention, even with a comparison of observations 1 and 3. Therefore, you can only guess whether the intervention caused any differences in the groups at time 2.

Despite this, the static-group comparison design is the best one for evaluating natural experiments, where you have no control over the assignment of participants anyway. The relation between smoking cigarettes (the intervention) and getting lung cancer (the dependent variable), for example, is easily seen by applying the humble *ex post facto* design with a control group for a second posttest.

In 1965, when the American Cancer Society (ACS) did its first big Cancer Prevention Study, men who smoked (that is, those who were subject to the intervention) were about 12 times more likely than nonsmokers (the control group) to die of lung cancer. At that time, relatively few women smoked and those who did had not been smoking very long. Their risk

was just 2.7 times that for women nonsmokers of dying from lung cancer.

Once those first ACS data were gathered, however, they became the baseline for a later study and by 1988, things had changed dramatically. Male smokers were then about 23 times more likely than nonsmokers to die of lung cancer, and female smokers were 12.8 times more likely than female nonsmokers to die of lung cancer. Men's risk had doubled (from about 12 to about 23), but women's risk had more than quadrupled (from 2.7 to about 13) (National Cancer Institute 1997).

The death rate for lung cancer has continued to fall among men in the United States, while the death rate for women has increased (http://apps.nccd.cdc.gov/uscs/) (Box 4.3).

Box 4.3 Migration and gender roles: A static-group comparison design

Lambros Comitas and I wanted to find out if the experience abroad of Greek labor migrants had any influence on men's and women's attitudes toward gender roles when they returned to Greece. The best design would have been to survey a group before they went abroad, then again while they were away, and again when they returned to Greece. Since this was not possible, we studied one group of persons who had been abroad and another group of persons who had never left Greece. We treated these two groups as if they were part of a static-group comparison design (Bernard and Comitas 1978).

From a series of life histories with migrants and nonmigrants, we learned that the custom of giving dowry was under severe stress (Bernard and Ashton-Vouyoucalos 1976). Our survey confirmed this: Those who had worked abroad were far less enthusiastic about providing expensive dowries for their daughters than were those who had never left Greece. We concluded that this was in some measure due to the experiences of migrants in West Germany.

There were threats to the validity of this conclusion: Perhaps migrants were a self-selected bunch of people who held the dowry and other traditional Greek customs in low esteem to begin with. But we had those life histories to back up our conclusion. Surveys are weak compared to true experiments, but their power is improved if they are conceptualized in terms of testing natural experiments and if their results are backed up with data from open-ended interviews.

The Interrupted Time Series Design

The interrupted time series design, shown in Figure 4.1h, can be very persuasive. It involves getting data from a series of points before and after an intervention and evaluating statistically whether the intervention has had an impact.

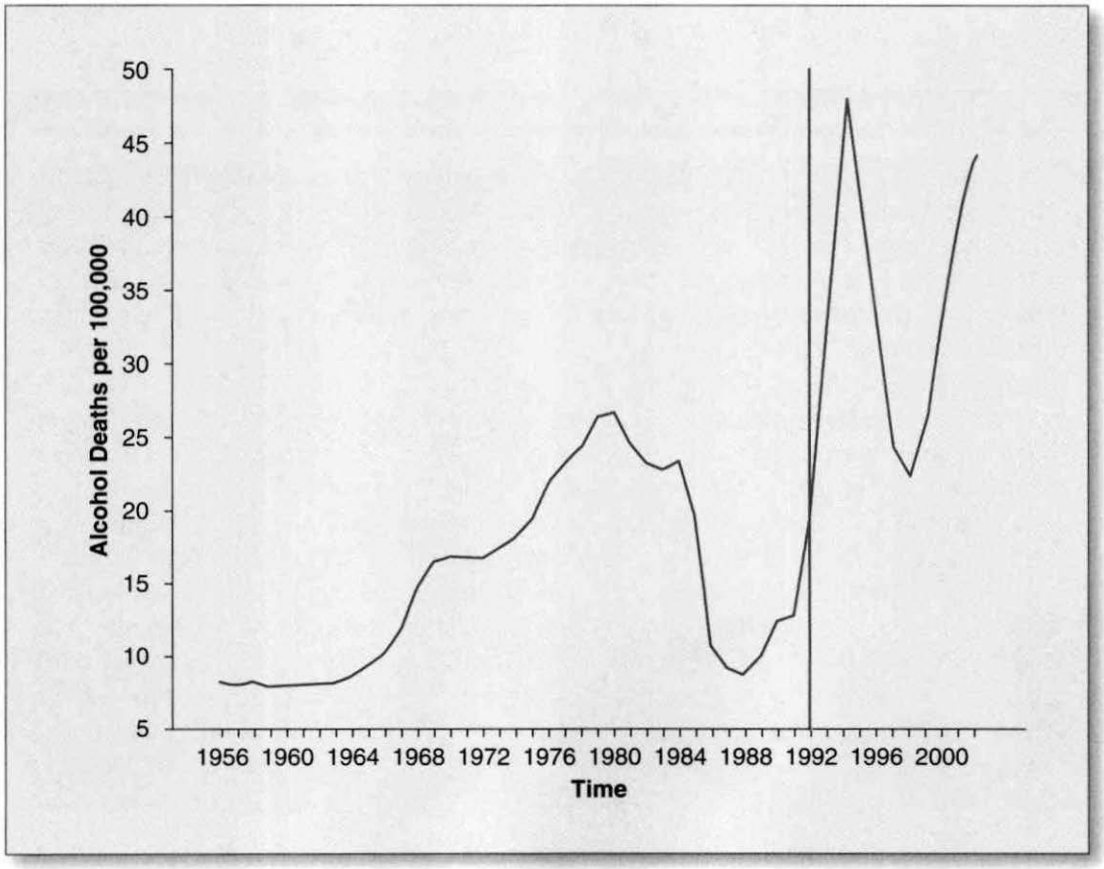
Figure 4.3 shows the rate of alcohol deaths, per *100,000* population, in Russia, from 1956 to 2002 (Pridemore et al. 2007:281). The year 1992 marks the formal shift in Russia to a market economy from a centrally planned one. As it turns out, homicide and suicide rates show very similar patterns—something that Pridemore et al. interpret as outcomes that are predictable from Durkheim's (1933 [1893], 1951 [1897]) theory of anomie.

The interrupted time series design is used to assess the effect of new laws. Until 1985, if two working spouses in Canada were married in January, then each of them paid full taxes on their own income. But if they married in December, one of them could claim the other as a dependent. The loophole was closed in

Figure 4.1h The Interrupted Time Series Design

	Time 1		Time 2	
	Assignment	Pretest	Intervention	Posttest
		OOO	*X*	*OOO*

Figure 4.3 Alcohol Deaths in Russia, 1956–2002: An Interrupted Time Series



Source: W. A. Pridemore et al., "An Interrupted Time-Series Analysis of Durkeim's Social Deregulation Thesis: The Case of the Russian Federation." *Justice Quarterly* 24:272–90, p. 281, 2007.

1986. Gelardi (1996) treated the new law as an interruption in a time series and found, just as we'd predict, a significant decrease in the percentage of marriages in December immediately after the new law took effect.

Gelardi found the same result when he examined times series data from England and Wales, where a similar change in the law had occurred in 1968. Bonham et al. (1992) found that Hawaii's 1987 tax of 5.25% on hotel rooms had no effect on rentals. In that case, the state simply made money and the hotel industry didn't suffer any loss.

THOUGHT EXPERIMENTS

As you can see, it is next to impossible to eliminate threats to validity in natural experiments.

However, there is a way to understand those threats and to keep them as low as possible: Think about research questions as if it were possible to test them in *true* experiments. These are called thought experiments.

This wonderful device is part of everyday culture in the physical sciences. In 1972, I did an ethnographic study of scientists at Scripps Institution of Oceanography. Here's a snippet from a conversation I heard among some physicists there. "If we could only get rid of clouds, we could capture more of the sun's energy to run stuff on Earth," one person said. "Well," said another, "there are no clouds above the Earth's atmosphere. The sun's energy would be lots easier to capture out there."

"Yeah," said the first, "so suppose we send up a satellite, with solar panels to convert sunlight to electricity, and we attach a really long extension cord so the satellite was tethered to the Earth. Would that work?" The discussion got weirder from there, if you can imagine, but it led to a lot of really useful ideas for research.

Suppose you wanted to know if Americans who own hand guns are more likely to get shot than are Americans who don't own hand guns. The experiment you'd have to set up is pretty macabre, but do the thought experiment nonetheless (no ethical issues are at stake in thinking). What experimental conditions would be required for you to be sure that both owning a hand gun and having a high probability of getting shot were not caused by some third factor, like place of residence and exposure to violent crime?

Or suppose you wanted to know if small farms can produce organically grown food on a scale sufficiently large to be profitable. What would a true experiment to test this question look like? You might select some smallish farms with similar acreage and assign half of them randomly to grow vegetables organically. You'd assign the other half of the farms to grow the same vegetables using all the usual technology (pesticides, fungicides, chemical fertilizers, and so on). Then, after a while, you'd measure some things about the farms' productivity and profitability and see which of them did better.

How could you be sure that organic or nonorganic methods of farming made the difference in profitability? Perhaps you'd need to control for access to the kinds of market populations that are friendly toward organically produced food (like university towns) or for differences in the characteristics of soils and weather patterns. Obviously, you can't do a true experiment on this topic, randomly assigning farmers to use organic or high-tech methods, but you *can* evaluate the experiments that real farmers are conducting every day in their choice of farming practices.

So, after you've itemized the possible threats to validity in your thought experiment, go out and look for natural experiments—societies, voluntary associations, organizations—that conform most closely to your ideal experiment. Then evaluate those natural experiments.

That's what Karen Davis and Susan Weller (1999) did in their study of the efficacy of condoms in preventing the transmission of HIV among heterosexuals. Here's the experiment you'd have to conduct. First, get 1,000 heterosexual couples. Make each couple randomly serodiscordant. That is, for each couple, randomly assign the man or the woman to be HIV-positive. Assign each couple randomly to one of three conditions: (1) they use condoms for each sexual act; (2) they sometimes use condoms; or (3) they don't use condoms at all. Let the experiment run a few years. Then see how many of the couples in which condoms are always used remain serodiscordant and how many become seroconcordant—that is, they are both HIV-positive. Compare across conditions and see how much difference it makes to always use a condom.

Clearly, no one could conduct such an experiment. But Davis and Weller scoured the literature on condom efficacy and found 25 studies that met three criteria: (1) the focus was on serodiscordant heterosexual couples who said they regularly had penetrative sexual

intercourse; (2) the HIV status of the subjects in each study had been determined by a blood test; and (3) there was information on the use of condoms. The 25 studies involved thousands of subjects, and from this meta-analysis Davis and Weller established that consistent use of condoms reduced the rate of HIV transmission by over 85% (**Further Reading:** thought experiments).

TRUE EXPERIMENTS IN THE LAB

Laboratory experiments to test theories about how things work in the real world is the preeminent method in social psychology. It has long been observed that fraternity hazing is difficult, dangerous, and painful—and produces people who come out of it supporting their tormentors. Remember Festinger's cognitive dissonance theory? That theory predicts that people who come out of a tough initiation experience (marine recruits at boot camp, prisoners of war, girls and boys who go through genital mutilation, etc.) wind up as supporters of their tormentors.

In a classic experiment, Elliot Aronson and Judson Mills (1959) recruited 63 college women for a discussion group that ostensibly was being formed to talk about psychological aspects of sex. To make sure that only mature people—people who could discuss sex openly—would make it into this group, some of the women would have to go through a screening test. Or at least that's what they were *told*.

A third of the women were assigned randomly to a group that had to read a list of obscene words and some sexually explicit passages from some novels—aloud, in front of a man who was running the experiment. (It may be hard to imagine now, but those women who went through this in the 1950s must have been very uncomfortable.) Another third were assigned randomly to a group that had to recite some nonobscene words that had to do with sex, and a third group went through no screening at all.

Then, each participant listened in on a discussion that was supposedly going on among the members of the group she was joining. The "discussion" was actually a recording and it was, as Aronson and Mills said, "one of the most worthless and uninteresting discussions imaginable" (Aronson and Mills 1959:179). The women rated the discussion, on a scale of 0–15, on things like dull-interesting, intelligent-unintelligent, and so on.

Those in the tough initiation condition rated the discussion higher than did the women in either the control group or the mild initiation group. Since all the women were assigned randomly to participate in one of the groups, the outcome was unlikely to have occurred by chance. The women in the tough initiation condition had gone through a lot to join the discussion. When they discovered how boringly nonprurient it was, what did they do? They convinced themselves that the group was worth joining.

Aronson and Mills's findings were corroborated by Gerard and Mathewson (1966) in an independent experiment. Those findings from the laboratory can now be the basis for a field test, across cultures, of the original hypothesis.

Conversely, events in the real world can stimulate laboratory experiments. In 1963, in Queens, New York, Kitty Genovese was stabbed to death in the street one night. There were reported to be 38 eye-witnesses who saw the whole grisly episode from their apartment windows, and not one of them called the police. The newspapers called it "apathy," but Bibb Latané and John Darley had a different explanation. They called it **diffusion of responsibility** and they did an experiment to test their idea (1968).

Latané and Darley invited ordinary people to participate in a "psychology experiment." While the subjects were waiting in an anteroom to be called for the experiment, the room filled with smoke. If there was a single subject in the room, 75% reported the smoke right

away. If there were three or more subjects waiting together, they reported the smoke only 38% of the time. People in groups just couldn't figure out whose responsibility it was to do something. So they did nothing.

As it turns out, there probably weren't 38 witnesses to the murder; none of the witnesses could have seen the whole episode; and some of the witnesses did call the police (Manning et al. 2007). Nevertheless, hundreds of studies on what's known as the bystander effect have been published since Latané and Darley did their pioneering work. A real event gave two scientists an idea that they tested in an experiment, and that experiment opened a whole area of research (**Further Reading:** diffusion of responsibility).

TRUE EXPERIMENTS IN THE FIELD

When experiments are done outside the lab, they are called field experiments. Several researchers have been studying the effect of being touched on consumers' spending. In one experiment (Hornik 1992), as lone shoppers (no couples) entered a large bookstore, an "employee" came up and handed them a catalog. Alternating between customers, the employee-experimenter touched about half the shoppers lightly on the upper arm.

The results? Across 286 shoppers, those who were touched spent an average of $15.03; those who were not touched spent just $12.23. (That's $26 vs. $21 in 2011 dollars.) The difference was across the board, no matter what the sex of the toucher or the shopper.

In another of his experiments, Hornik enlisted the help of eight servers—four men and four women—at a large restaurant. At the end of the meal, the servers asked each of 248 couples (men and women) how the meal was. Right then, for half the couples, the servers touched the arm of either the male or the female in the couple for one second. The servers didn't know it, but they had been selected out of 27 servers in the restaurant to represent two ends of a physical attractiveness scale. The results? Men and women alike left bigger tips when they were touched, but the effect was stronger for women patrons than for men. Overall, couples left about a 19% tip when the woman was touched, but only 16.5% when the man was touched. (See Seiter [2007] for more field experiments on tipping.)

Ronald Milliman (1986) tested the effects of slow and fast music on the behavior of customers in a restaurant. First, Milliman played a number of instrumental pieces as background music. He asked 227 randomly chosen customers "Do you consider the music playing right now as slow tempo, fast tempo, or in between?" From these data, he identified slow music as 72 beats per minute or fewer and fast music as 92 beats per minute or more.

Then, for eight consecutive weekends, Milliman played slow music and fast music on alternating nights. The first weekend, he played slow music on Friday night and fast music on Saturday night. The next weekend he reversed the order—just in case different kinds of people like to go out on Friday and Saturday night. This procedure simulated assigning customers randomly to the different conditions of slow or fast music. Milliman used only instrumental music in order not to confound the experiment with exogenous variables like gender of vocalist, popularity of vocalist, and so on.

Milliman looked at the effect of the two music tempos on six dependent variables. Music tempo had no effect on five of those variables. It had no effect on the time it took for employees to take, prepare, and serve customers' orders. It had no effect on the number of people who decided to leave the restaurant before being seated (it was a popular restaurant, and there was usually a wait on weekends for a table). And it had no effect on the total dollar amount of food purchased.

Music tempo had a significant effect, however, on the amount of time that customers

spent at their tables. With slow music, customers spent 56 minutes eating; with fast music, they spent only 45 minutes. For tables that were occupied those extra 11 minutes, the average bar tab was $30.47, about $9 more than the average bar tab per table in the fast-music treatment. Since the amount of food purchased was the same under both conditions, and since profits are much higher on bar purchases, the total profit margin per table was dramatically higher in the slow music treatment.

Marvin Harris and his colleagues (1993) conducted a field experiment in Brazil to test the effect of substituting one word in the question that deals with race on the Brazilian census. The demographers who designed the census had decided that the term *parda* was a more reliable gloss than *morena* for what English speakers call "brown," despite overwhelming evidence that Brazilians prefer the term *morena*.

In the town of Rio de Contas, Harris et al. assigned 505 houses randomly to one of two groups and interviewed one adult in each house. All respondents were asked to say what *cor* (color) they thought they were. This was the free-choice option. Then they were asked to choose one of four terms that best described their *cor*. One group (with 252 respondents) was asked to select among *branca* (white), *parda* (brown), *preta* (black), and *amerela* (yellow). This was the "*parda* option"—the one used on the Brazilian census. The other group (with 253 respondents) was asked to select among *branca*, *morena* (brown), *preta*, and *amerela*. This was the "*morena* option," and is the intervention, or treatment in Harris's experiment.

Among the 252 people given the *parda* option, 131 (52%) identified themselves as *morena* in the free-choice option (when simply asked to say what color they were). But when given the *parda* option, only 80 of those people said they were *parda* and 41 said they were *branca* (the rest chose the other two categories). Presumably, those 41 people would have labeled themselves *morena* if they'd had the chance; not wanting to be labeled *parda*, they said they were *branca*. The *parda* option, then, produces more Whites (*brancas*) in the Brazilian census and fewer Browns (*pardas*).

Of the 253 people who responded to the *morena* option, 160 (63%) said they were *morena*. Of those 160, only 122 had chosen to call themselves *morena* in the free-choice option. So, giving people the *morena* option actually increases the number of Browns (*morenas*) and decreases the number of Whites (*brancas*) in the Brazilian census.

Does this difference make a difference? Social scientists who study the Brazilian census have found that those who are labeled Whites live about seven years longer than do those labeled non-Whites in that country. If 31% of self-described *morenas* say they are Whites when there is no *morena* label on a survey and are forced to label themselves *parda*, what does this do to all the social and economic statistics about racial groups in Brazil? (Harris et al. 1993) (**Further Research:** field experiments).

NATURAL EXPERIMENTS

True experiments and quasi-experiments are *conducted* and the results are *evaluated* later. Natural experiments, by contrast, are going on around us all the time. They are not conducted by researchers at all—they are simply evaluated.

Here are four examples of common natural experiments: (1) Some employees in a company get expanded job responsibilities; some do not. (2) Some young people choose to migrate from small, isolated communities in northern Quebec to Montreal; others stay put. (3) Some second-generation, middle-class Mexican American students go to college; some do not. (4) Some cultures practice female infanticide; some do not.

Each of these situations constitutes a natural experiment that tests *something* about human behavior and thought. The trick is to

ask "What hypothesis is being tested by what's going on here?"

To evaluate natural experiments—that is, to figure out what hypothesis is being tested—you need to be alert to the possibilities and collect the right data. There's a really important natural experiment going in an area of Mexico where I've worked over the years. A major irrigation system has been installed over the last 50 years in parts of a desert valley. Some of the villages affected by the irrigation system are populated entirely by Ñähñu (Otomí) Indians; other villages are entirely mestizo (as the majority population of Mexico is called).

Some of the Indian villages in the area are too high up the valley slope for the irrigation system to reach. I could not have decided to run this multimillion dollar system through certain villages and bypass others, but the instant the decision was made by others, a natural experiment on the effects of a particular intervention was set in motion. There is a treatment (irrigation), there are treatment groups (villages full of people who get the irrigation), and there are control groups (villages full of people who are left out).

Unfortunately, I can't evaluate the experiment because I simply failed to see the possibilities early enough. Several studies of the region show that the intervention is having profound effects, but no one thought to measure things at the village level like average wealth or rates of migration, alcoholism, literacy, diabetes . . . things that could easily have been affected by the coming of irrigation. Had anyone done so—if we had baseline data—we would be in a better position to ask "What hypotheses about human behavior are being tested by this experiment?" I can't reconstruct variables from 50 years ago. The logical power of the experimental model for establishing cause and effect between the intervention and the dependent variables is destroyed.

Some natural experiments, though, like the famous 1955 Connecticut speeding law, produce terrific data all by themselves for evaluation. In 1955, the governor of Connecticut ordered strict enforcement of speeding laws in the state. The object was to cut down on the alarming number of traffic fatalities. Anyone caught speeding had their driver's license suspended for at least 30 days. Traffic deaths fell from 324 in 1955 to 284 in 1956. A lot of people had been inconvenienced with speeding tickets and suspension of driving privileges, but 40 lives had been saved.

The question was whether the crackdown was the cause of the decline in traffic deaths. Campbell and Ross (1968) used the available data to find out. They plotted the traffic deaths for 1951 to 1959 in Connecticut, Massachusetts, New York, New Jersey, and Rhode Island. Each of those states has more or less the same weather, and they all produced good data on traffic fatalities. Four of the five states showed an increase in highway deaths in 1955, and all five states showed a decline in traffic deaths the following year, 1956. If that were all you knew, you couldn't be sure about the cause of the decline. However, traffic deaths continued to decline steadily in Connecticut for the next three years (1957, 1958, 1959). In Rhode Island and Massachusetts, they went up; in New Jersey, they went down a bit and then up again; and in New York, they remained about the same.

Connecticut was the only state that showed a consistent reduction in highway deaths for four years after the stiff penalties were introduced. Campbell and Ross treated these data as a series of natural experiments, and the results were convincing: Stiff penalties for speeders saves lives.

Natural Experiments Are Everywhere

If you think like an experimentalist, you eventually come to see the unlimited possibilities for research going on all around you. Across the United States, school boards set rigid cutoff dates for children who are starting school. Suppose the cutoff is August 1. Children born

at the end of July start kindergarten at age 5. Children born at the beginning of August start at age 6. Since children from 5 to 7 are going through an intense period of cognitive development, Morrison et al. (1996) treat this situation as a natural experiment and ask: Are there short- and long-term impacts on cognitive skills of just missing or just making the cutoff?

Ever notice how people like to tell stories about the time they found themselves sitting next to a famous person on a plane? It's called BIRGing in social psychology—basking in reflected glory. Cialdini et al. (1976) evaluated the natural BIRGing experiment that is conducted on most big university campuses every weekend during football season. Over a period of eight weeks, professors at Arizona State, Louisiana State, Ohio State, Notre Dame, Michigan, the University of Pittsburgh, and the University of Southern California recorded the percentage of students in their introductory psychology classes who wore school insignias (buttons, hats, t-shirts, etc.) on the Monday after Saturday football games.

For 177 students per week, on average, over eight weeks, 63% wore some school insignia after wins in football versus 44% after losses or ties. The difference was statistically significant and the finding opened up a whole area of research that continues (Madrigal and Chen 2008).

Here's another one. On January 1, 2002, 12 of the then-15 members of the European Union gave up their individual currencies and adopted the euro (there are now 17 countries in the eurozone out of 27 in the European Union). Greece was one of the 12, Denmark wasn't. Some researchers noticed that many of the euro coins were smaller than the Greek drachma coins they'd replaced and thought that this might create a choking hazard for small children (Papadopoulos et al. 2004). The researchers compared the number of choking incidents reported in Danish and Greek hospitals in January through March from 1996 through 2002.

Sure enough, there was no increase in the rate of those incidents in Denmark (which hadn't converted to the euro), but the rate in Greece suddenly more than doubled in 2002 (Box 4.4).

Box 4.4 Case control and natural experiments

In a **case control design**, you compare naturally occurring cases of a criterion (like having a certain illness or injury, or attempting suicide, or being homeless) with people who match the cases on many criteria, but *not* on the case criterion.

This method is widely used in public health research. The first link between cigarette smoking and lung cancer, for example, was the result of a case-control study (Wynder and Graham 1950). A classic case-control study in the social sciences was done by Art Rubel and colleagues (1984) on the Latin American folk illness known as *susto*. Among Indian people, the symptoms of *susto*—anxiety, diarrhea, difficulty breathing, and others—are often said to result from a loss of soul, while among mestizo people the symptoms are often attributed to having experienced some fright (from the verb *asustarse*, to become frightened).

Rubel and his colleagues compared people in two Indian villages and one mestizo village in Mexico, all of whom suffered from *susto* (the **index cases**) with people in the same villages who did not (the **control cases**). There were no statistically significant differences between the 47 index cases and the 48 controls on tests of psychiatric symptoms. Seven years after the study was completed, however, 17% of the index cases had died—and *none* of the control cases had died. The design in this study makes it thoroughly convincing.

There is a well-known hypothesis in psychology called the "goal-gradient hypothesis" or "deadline hypothesis." In 1934, Clark Hull showed that the closer rats came to food, the faster they went. Since then, we've learned that the number of plays in a football game is highest in the second quarter, next highest in the fourth quarter, and lowest in the first and third quarters; trading goes up in the last two hours of the day at the New York Stock Exchange (Webb and Weick 1983); and when people get one of those "Buy ten coffees, get one free" cards, they buy coffees more frequently, the closer they get to the reward (Kivetz et al. 2006). Apparently, people, like rats, perform the most when they face a deadline.

Data for evaluating natural experiments can come from direct observation (counting up the t-shirts with school colors) or from archives (records of stock trades), but they can also come from survey questionnaires. Emotional pressures are among the predictors of alcoholism. Before 1975, all men 18 years old and older in the United States were subject to the military draft—a kind of lottery that determined who would spend at least two years in the armed services. Goldberg et al. (1991) reasoned that men who had been subject to the military draft were subject to more emotional pressures than those who weren't. Goldberg et al. evaluated this natural experiment with data on over 1,800 male respondents in the National Health Interview Surveys of 1977, 1983, and 1985.

As it turned out, men who had been eligible for the draft were *not* more likely to be heavy consumers of alcohol later in life than were men who had never been eligible for the draft. But the men who had been eligible for the draft were more likely to have served in the military, and those men were more likely than men who had not served to be heavy drinkers.

NATURALISTIC EXPERIMENTS

In a naturalistic experiment, you contrive to collect experimental data under natural conditions. You make the data happen out in the natural world (not in the lab) and you evaluate the results.

In a memorable experiment, elegant in its simplicity of design, Doob and Gross (1968) had a car stop at a red light and wait for 15 seconds after the light turned green before moving again. In one experimental condition, they used a new car and a well-dressed driver. In another condition, they used an old, beat-up car and a shabbily dressed driver. They repeated the experiment many times and measured the time it took for people in the car behind the experimental car to start honking their horns. It won't surprise you to learn that people were quicker to vent their frustration at apparently low-status cars and drivers.

Piliavin et al. (1969) did a famous naturalistic experiment to test the "good Samaritan" problem. Students in New York City rode a particular subway train that had a 7.5-minute run at one point. At 70 seconds into the run, a researcher pitched forward and collapsed. The team used four experimental conditions: The "stricken" person was either Black or White and was either carrying a cane or a liquor bottle. Observers noted how long it took for people in the subway car to come to the aid of the supposedly stricken person, the total population of the car, whether bystanders were Black or White, and so on. You can conjure up the results. There were no surprises.

Harari et al. (1985) recruited drama majors to test whether men on a college campus would come to the aid of a woman being raped. They staged realistic-sounding rape scenes and found that there was a significant difference in the helping reaction of male passersby if those men were alone or in groups (**Further Reading:** naturalistic experiments).

The Small-World Experiment

Consider this: You're having coffee near the Trevi Fountain in Rome. You overhear two Americans chatting next to you and you ask where they're from. One of them says he's from Sioux City, Iowa. You say you've got a

friend from Sioux City and it turns out to be your new acquaintance's cousin. The culturally appropriate reaction at this point is for everyone to say, "Wow, what a small world."

Stanley Milgram (1967) contrived an experiment to test how small the world really is. He asked a group of people in the midwestern United States to send a folder to a divinity student at Harvard University, but only if the subject *knew* the divinity student personally. Otherwise, he asked them to send the folders to an acquaintance whom they thought had a chance of knowing the "target" at Harvard.

The folders got sent around from acquaintance to acquaintance until they wound up in the hands of someone who actually knew the target—at which point the folders were sent, as per the instructions in the game, to the target. The average number of links between all the "starters" and the target was about five. It really *is* a small world.

No one expects this experiment to actually happen in real life. It's contrived as can be and lacks control. On the other hand, it's compelling because it says *something* about how the natural world works. The finding was so compelling that it was the basis for the Broadway play *Six Degrees of Separation*, as well as the movie of the same name that followed and the game *Six Degrees of Kevin Bacon*. It also provoked research, first by social scientists and then by physicists and mathematicians, on the structure of relations in everything from people to corporations to nations to neurons to sites on the Internet. Looking for the structure of relations is another way of saying network analysis (**Further reading:** small-world research).

The Lost-Letter Technique

Another of Milgram's contributions is a method for doing unobtrusive surveys of political opinion. The method is called the "lost-letter technique" and consists of "losing" a lot of letters that have addresses and stamps on them (Milgram et al. 1965).

The technique is based on two assumptions. First, people in many societies believe that they ought to mail a letter if they find one, especially if it has a stamp on it. Second, people will be less likely to drop a lost letter in the mail if it is addressed to someone or some organization that they don't like.

Milgram et al. (1965) tested this in an experiment in New Haven, Connecticut. They lost 400 letters in 10 districts of the city. They dropped the letters on the street; they left them in phone booths; they left them on counters at shops; and they tucked them under windshield wipers (after penciling "found near car" on the back of the envelope). Over 70% of the letters addressed to an individual or to a medical research company were returned. Only 25% of the letters addressed to either "Friends of the Communist Party" or "Friends of the Nazi Party" were returned. (The addresses were all the same post box that had been rented for the experiment.)

By losing letters in a sample of communities, and by calculating the different rates at which they are returned, you can test variations in sentiment. Two of Milgram's students distributed anti-Nazi letters in Munich. The letters did not come back as much from some neighborhoods as from others, and they were thus able to pinpoint the areas of strongest neo-Nazi sentiment (Milgram 1969:68).

Bushman and Bonacci (2004) extended the lost-letter technique to e-mail by purposely sending messages out that appeared to be intended for someone other than the recipient. The surname of the supposed recipient was clearly either European or Arabic. The message said that the recipient either had or had not been awarded a prestigious four-year scholarship and that a response was required within 48 hours. The recipients of these messages were 512 intro psych college students who had filled out a questionnaire two weeks prior to the experiment, so the researchers had a lot of information about those recipients, including scores on a scale of prejudice against

various minorities (Asians, Hispanics, African Americans, and Arab Americans). Those with higher prejudice scores were less likely to return the message with good news for an Arab and more likely to return the message with bad news.

The lost-letter technique has sampling problems and validity problems galore associated with it. But you can see just how intuitively powerful the results can be (**Further Reading:** the lost-letter technique) (Box 4.5).

Box 4.5 Naturalistic experiments don't have to be complicated

Walker (2006) rode his bicycle 200 miles through Bristol and Salisbury England during regular working hours dressed as an ordinary commuter–sometimes wearing a helmet, sometimes not; sometimes wearing a woman's long-haired wig, sometimes not. Walker outfitted his bike with a hidden distance sensor and a tiny camera and then, systematically varying his distance from the curb, he measured how close cars came as they passed him. The further from the edge he rode, the closer drivers came; drivers stayed further from him when he appeared to be a woman; and drivers came closer to him when he wasn't wearing a helmet than when he was. Fortunately, both times he got hit doing this experiment, he was wearing the helmet.

Comparative Field Experiments

Naturalistic field experiments appeal to me because they are excellent for comparative research, and comparison is so important for developing theory. Feldman (1968) did five field experiments in Paris, Boston, and Athens to test whether people in those cities respond more kindly to foreigners or to members of their own culture.

In one experiment, the researchers simply asked for directions and measured whether foreigners or natives got better treatment. Parisians and Athenians gave help significantly more often to fellow citizens than to foreigners. In Boston, there was no difference.

In the second experiment, foreigners and natives stood at major metro stops and asked total strangers to do them a favor. They explained that they were waiting for a friend, couldn't leave the spot they were on, and had to mail a letter. They asked people to mail the letters for them (the letters were addressed to the experiment headquarters) and simply counted how many letters they got back from the different metro stops in each city. Half the letters were unstamped.

In Boston and Paris, between 32% and 35% of the people refused to mail a letter for a fellow citizen. In Athens, 93% refused. Parisians treated Americans significantly better than Bostonians treated Frenchmen on this task. In fact, in cases where Parisians were asked to mail a letter that was stamped, they treated Americans significantly better than they treated other Parisians. (So much for *that* stereotype.)

In the third experiment, researchers approached informants and said: "Excuse me, sir. Did you just drop this dollar bill?" (or other currency, depending on the city). It was easy to measure whether or not people falsely claimed the money more from foreigners than from natives. This experiment yielded meager results.

In the fourth experiment, foreigners and natives went to pastry shops in the three cities, bought a small item, and gave the clerk 25% more than the item cost. Then they left the shop and recorded whether the clerk had offered to return the overpayment. This experiment also showed little difference among the

cities or between the way foreigners and locals are treated.

And in the fifth experiment, researchers took taxis from the same beginning points to the same destinations in all three cities. They measured whether foreigners or natives were charged more. In neither Boston nor Athens was a foreigner overcharged more than a local. In Paris, however, Feldman found that "the American foreigner was overcharged significantly more often than the French compatriot in a variety of ingenious ways" (1968:11).

Feldman collected data on more than 3,000 interactions and was able to draw conclusions about cultural differences in how various peoples respond to foreigners as opposed to other natives. Some stereotypes were confirmed; others were crushed. Since Feldman's pioneering work, dozens of studies have been done on cross-cultural differences in helping strangers (see Levine et al. [2001], for example).

Bochner's Field Experiments in Australia

Bochner did a series of interesting experiments on the nature of Aboriginal-White relations in urban Australia (see Bochner [1980:335–40] for a review). These experiments are clever, inexpensive, and illuminating, and Bochner's self-conscious critique of the limitations of his own work is a model for field experimentalists to follow. In one experiment, Bochner put two classified ads in a Sydney paper:

> Young couple, no children, want to rent small unfurnished flat up to $25 per week. Saturday only. 759–6000.
>
> Young Aboriginal couple, no children, want to rent small unfurnished flat up to $25 per week. Saturday only. 759–6161. [Bochner 1972:335]

Different people were assigned to answer the two phones, to ensure that callers who responded to both ads would not hear the same voice. Note that the ads were identical in every respect, except for fact that in one of the ads the ethnicity of the couple was identified and in the other it was not. There were 14 responses to the ethnically nonspecific ad and two responses to the ethnically specific ad (three additional people responded to both ads).

In another experiment, Bochner exploited what he calls the "Fifi effect" (Bochner 1980:336). The Fifi effect refers to the fact that urbanites acknowledge the presence of strangers who pass by while walking a dog and ignore others. Bochner sent a White woman and an Aboriginal woman, both in their early 20s, and similarly dressed, to a public park in Sydney. He had them walk a small dog through randomly assigned sectors of the park, for 10 minutes in each sector.

Each woman was followed by two observers, who gave the impression that they were just out for a stroll. The two observers *independently* recorded the interaction of the women with passersby. The observers recorded the frequency of smiles offered to the women, the number of times anyone said anything to the women, and the number of nonverbal recognition nods the women received. The White woman received 50 approaches; the Aboriginal woman received only 18 (Bochner 1971:111).

There are many elegant touches in this experiment. Note how the age and dress of the experimenters were controlled so that only their ethnic identity remained as a dependent variable. Note how the time for each experimental trial (10 minutes in each sector) was controlled to ensure an equal opportunity for each woman to receive the same treatment by strangers. Bochner did preliminary observation in the park and divided it into sectors that had the same population density so that the chance for interaction with strangers would be about equal in each run of the experiment, and he used two independent observer-recorders.

As Bochner points out, however, there were still design flaws that threatened the internal validity of the experiment (1980:337). As it happens, the interrater reliability of the two observers in this experiment was nearly perfect. But suppose the two observers shared the same cultural expectations about Aboriginal-White

relations in urban Australia. They might have quite reliably misrecorded the cues that they were observing.

Reactive and unobtrusive observations alike tell you *what* happened, not *why*. It is tempting to conclude that the Aboriginal woman was ignored because of active prejudice. But, says Bochner, "perhaps passersby ignored the Aboriginal . . . because they felt a personal approach might be misconstrued as patronizing" (Bochner 1980:338).

In Bochner's third study, a young White or Aboriginal woman walked into a butcher's shop and asked for 10 cents' worth of bones for her pet dog (about 50 cents in current Australian dollars). The dependent variables in the experiment were the weight and quality of the bones. (An independent dog fancier rated the bones on a three-point scale, without knowing how the bones were obtained, or why.) Each woman visited seven shops in a single middle-class shopping district.

In both amount and quality of bones received, the White woman did better than the Aboriginal, but the differences were not statistically significant—the sample was just too small so no conclusions could be drawn from that study alone. *Taken all together*, though, the three studies done by Bochner and his students comprise a powerful set of information about Aboriginal-White relations in Sydney. Naturalistic experiments like these have their limitations, but they often produce intuitively compelling results. And since Bochner's work, over 30 years ago, dozens of other field studies have been done testing for discrimination in housing, lending, and hiring (Ahmed and Hammarstedt 2008; Sharpe 1998).

ARE FIELD EXPERIMENTS ETHICAL?

Field experiments come in a range of ethical varieties, from innocuous to borderline to downright ugly. I see no ethical problems with the lost-letter technique. When people mail one of the lost letters, they don't know that they are taking part in a social science experiment, but that doesn't bother me. Personally, I see no harm in the experiment to test whether people vent their anger by honking their car horns more quickly at people they think are lower socioeconomic class. These days, however, with road rage an increasing problem, I do not recommend repeating Doob and Gross's experiment.

Randomized field experiments, used mostly in evaluation research, can be problematic. Suppose you wanted to know whether fines or jail sentences are better at changing the behavior of drunk drivers. One way to do that would be to randomly assign people who were convicted of the offense to one or the other condition and watch the results. Suppose one of subjects whom you didn't put in jail kills an innocent person?

The classic experimental design in drug testing requires that some people get the new drug, that some people get a placebo (a sugar pill that has no effect), and that neither the patients nor the doctors administering the drugs know which is which. This double-blind placebo design is responsible for great advances in medicine and the saving of many lives. But suppose that, in the middle of a double-blind trial of a drug you find out that the drug really works. Do you press on and complete the study? Or do you stop right there and make sure that you aren't withholding treatment from people whose lives could be saved? The ethical problems associated with withholding of treatment are under increasing scrutiny (Storosum et al. 2003; Walther 2005; Wertz 1987).

There is a long history of debate about the ethics of deception in psychology and social psychology (see Hertwig and Ortmann [2008] for a review). My own view is that, on balance, some deception is clearly necessary—certain types of research just can't be done without it. When you use deception, though, you run all kinds of risks—not just to research subjects, but to the research itself. These days, college

students (who are the subjects for most social psych experiments) are very savvy about all this and are on the lookout for clues as to the "real" reason for an experiment the minute they walk in the door.

If you don't absolutely need deception in true behavioral experiments, that's one less problem you have to deal with. If you decide that deception is required, then understand that the responsibility for any bad outcomes is yours and yours alone.

The experiments by Piliavin et al. (1969) and Harari et al. (1985) on whether people will come to the aid of a stricken person or a woman being raped are real ethical problems. Some of the participants (who neither volunteered to be in an experiment nor were paid for their services) might still be wondering what happened to that poor guy on the subway whom they stepped over in their hurry to get away from an uncomfortable situation—or that woman whose screams they ignored. In laboratory experiments, at least, subjects are debriefed—told what the real purpose of the study was—to reduce emotional distress. In the guerrilla theater type of field experiment, though, no debriefing is possible (Box 4.6).

Box 4.6 Even debriefing has its dark side

Experiments with human subjects, both laboratory and field experiments, pose threats of harm that survey research or ethnographic research don't share. A great many experiments, as we have seen, involve deception–fooling subjects into thinking that an experiment is about one thing when it is really about something else entirely. Research subjects in these experiments may experience a loss of self-esteem when they find out that they've been duped.

If an experiment requires deception, make sure a debriefing session is a routine part of the design. At the debriefing, the true goal of the experiment is divulged. Subjects are told why deception was required, are thanked for their cooperation, and are given the opportunity to talk about any problems they may have with their participation in the experiment. The idea is to help people feel good about their cooperation in a scientific experiment rather than foolish about their having been hoodwinked. Some people may experience a loss of self-esteem if they conclude that they acted badly in an experiment.

How would *you* feel if you were one of the people who failed to respond to a rape victim, and then were told that you were just part of an experiment–that no real rape ever took place, and thank you very much for your help? (**Further Reading:** deception and debriefing).

If you think some of *these* cases are borderline, consider the study by West et al. (1975) on whether there is a little larceny in us all.

The Watergate Experiment

In the Watergate affair of 1974, men loyal to then President Richard Nixon broke into the headquarters of the Democratic Party at the Watergate Hotel in Washington, DC, to photograph documents pertinent to the 1972 election campaign. Their bungling of the job, and the subsequent cover-up by Nixon and his staff at the White House, led to the unprecedented resignation of the president of the United States from office in 1974. Soon after, West et al. conducted their experiment.

They confronted 80 different students with a proposition to burglarize a local advertising firm. Subjects were randomly assigned to one

of four conditions. In the first condition, subjects were told that the job was to be done for the Internal Revenue Service. The IRS, it seemed, needed to get the goods on this company to bring them to trial for tax evasion. If the subjects were caught in the act, then the government would guarantee immunity from prosecution. In the second condition, subjects were told that there was no immunity from prosecution.

In the third condition, subjects were told that another advertising agency had paid $8,000 for the job, and that they (the subjects) would get $2,000 for their part in it. (Remember, that was $2,000 in 1975—about $9,000 today.) Finally, in the fourth condition, subjects were told that the burglary was being committed just to see if the plan would work. Nothing would be taken from the office.

Understand that this was not a "let's pretend" exercise. People were not brought into a laboratory and told to imagine that they were being asked to commit a crime. This was for real. Subjects met the experimenter at his home or at a restaurant. They were all criminology students at a university and knew the experimenter to be an actual local private investigator. The private eye arranged an elaborate and convincing plan for the burglary, including data on the comings and goings of police patrol cars, aerial photographs, blueprints of the building—the works.

The subjects of this experiment really believed that they were being solicited to commit a crime. Just as predicted by the researchers, a lot of them agreed to do it in the first condition, where they thought the crime was for a government agency and that they'd be free of danger from prosecution if caught. What do you suppose would happen to *your* sense of self-worth when you were finally debriefed and told that you were one of the 36 out of 80 (45%) who agreed to participate in the burglary in the first condition? (See S. W. Cook [1975] for a critical comment on the ethics of this experiment.)

FACTORIAL DESIGNS: MAIN EFFECTS AND INTERACTION EFFECTS

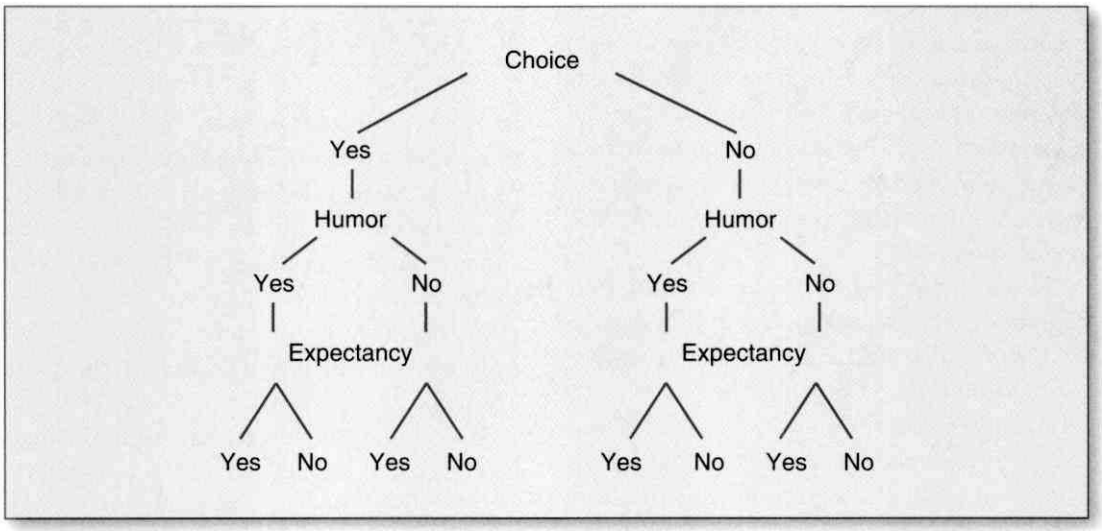
Most experiments involve analyzing the effects of several independent variables at once. A factorial design lays out all the combinations of all the categories of the independent variables. That way you know how many subjects you need, how many to assign to each condition, and how to run the analysis when the data are in.

It is widely believed that a good laugh has healing power. Rotton and Shats (1996) developed an experimental design to test this. They recruited 39 men and 39 women who were scheduled for orthopedic surgery. The patients were assigned randomly to one of nine groups—eight experimental groups and one control group. The patients in the eight treatment groups got to watch a movie in their room the day after their surgery.

There were three variables: choice, humor, and expectancy. The participants in the high-choice group got a list of 20 movies from which they chose four. The participants in the low-choice group watched a movie that one of the people in the high-choice group had selected. Half the subjects watched humorous movies, and half watched action or adventure movies. Before watching their movie, half the subjects read an article about the benefits of humor, while half read an article about the healthful benefits of exciting movies.

Figure 4.4 is a branching tree diagram that shows how these three variables, each with two attributes, create the eight logical groups for Rotton and Shats's experiment. Table 4.1 shows the same eight-group design, but in a format that is more common. The eight nodes at the bottom of the tree in Figure 4.4 and the sets of numbers in the eight boxes of

Figure 4.4 The Eight Conditions in Rotton and Shat's 2 × 2 × 2 Design



Source: J. Rotton and M. Shats, "Effects of State Humor, Expectancies, and Choice on Postsurgical Mood and Self-Medication: A Field Experiment," *Journal of Applied Social Psychology*, Vol. 26, pp.1775–94, 1996, Wiley-Blackwell.

Table 4.1 Three-Way, 2 × 2 × 2, Factorial Design

			Variable 3	
		Variable 2	Attribute 1	Attribute 2
Variable 1	Attribute 1	Attribute 1	1,1,1 Condition 1	1,1,2 Condition 2
		Attribute 2	1,2,1 Condition 3	1,2,2 Condition 4
	Attribute 2	Attribute 1	2,1,1 Condition 5	2,1,2 Condition 6
		Attribute 2	2,2,1 Condition 7	2,2,2 Condition 8

Table 4.1 are called the conditions in a factorial design.

The dependent variables in this study included a self-report by patients on the amount of pain they had and a direct measure of the amount of pain medication they took. All the patients had an access device that let them administer more or less of the analgesics that are used for controlling pain after orthopedic surgery.

In assessing the results of a factorial experiment, researchers look for main effects

and interaction effects. Main effects are the effects of each independent variable on each dependent variable. Interaction effects are effects on dependent variables that occur as a result of *interaction* between two or more independent variables. In this case, Rotton and Shats wanted to know the effects of humor on postoperative pain, but they wanted to know the effect *in different contexts*: in the context of choosing the vehicle of humor or not, in the context of being led to believe that humor has healing benefits or not, and so on.

As it turned out, being able to choose their own movie had no effect when patients saw action films. But patients who saw humorous films and who had not been able to make their own choice of film gave themselves more pain killer than did patients who saw humorous films and had been able to make the selection themselves (Rotton and Shats 1996).

We'll look at how to measure these effects when we take up ANOVA, or analysis of variance, in Chapter 21.

Key Concepts in This Chapter

logic of the experimental method
experimental thinking
research designs
threats to validity
randomized assignment
nonrandomized assignment
true experiments
quasi-experiments
treatment group
control group
field experiments
natural experiments
naturalistic experiments
classic experiment
confirmatory research
exploratory research
intervention group
stimulus group
experimental conditions
systematic bias
selection bias
pretest
posttest
confounds to validity
internal validity
cumulative knowledge
external validity
history confound
maturation confound
testing confound
instrumentation confound
interrater reliability
regression to the mean
mortality confound
diffusion of treatments
two-group pretest-posttest design
pretests
posttests
Solomon four-group design
interaction effect
two-group pretest-posttest design without random assignment
posttest-only design with random assignment
Campbell and Stanley posttest-only design
factorial design
one-shot case study
ex post facto design
one-group pretest-posttest design
two-group posttest only design without random assignment
static group comparison
interrupted time series design
thought experiments
meta-analysis
diffusion of responsibility
exogenous variables
diffusion of responsibility
baseline data
case control design
index cases
control cases
withholding of treatment
deception
debriefing
conditions in a factorial design
main effects
interaction effects

Summary

- The experimental method is not one single technique, but an approach to the development of knowledge. In experiments, researchers try to control the effects of confounds to understand the effects of particular independent variables on outcomes. In other words, they try to maximize the internal validity of experiments.
  - The well-known threats to validity include: history, maturation, testing, regression to the mean, selection bias, mortality of subjects (referring to the subjects of experiments dropping out), and diffusion of treatment.
- Different experimental designs control for various threats to validity. Among the widely used designs are: the two-group, pretest-posttest design with random assignment, the Solomon four-group design, the two-group pretest-posttest design without random assignment, the posttest only design with random assignment, the one-shot case study, the one-group pretest-posttest, two-group posttest only design (also called the static group comparison), and the interrupted time series.
  - Some designs are more effective than others, but it is not possible to use the most effective designs in all situations.
- You don't necessarily need a laboratory to carry out a true social science experiment. Many experiments are conducted in the field.
  - Natural experiments, quasi-experiments, and naturalistic experiments are based on field research rather than on laboratory research.
- Social and behavioral experiments raise particularly serious ethical problems. Even with proper debriefing, participants in experiments may experience loss of self-esteem. The need, or lack of need, for deception in social research is a matter of continuing debate.
- Most experiments involve analyzing the effects of several independent variables at once. A factorial design lays out all the combinations of all the categories of the independent variables. This sets up a blueprint for systematic data analysis.
  - In assessing the results of a factorial experiment, researchers look for main effects and interaction effects. Main effects are the effects of each independent variable on each dependent variable. Interaction effects are effects on dependent variables that occur as a result of interaction between two or more independent variables.

Exercises

1. For a random sample of students on your campus, count the number who are wearing school color—clothes, insignia, etc.—on Mondays following sports wins versus Mondays following sports losses. This replicates the study I discussed earlier by Cialdini et al. (1976). See Chapter 5 about taking a random sample.

2. Suppose you were designing an experiment to test whether a new diet plan, coupled with a motivational seminar, helped obese people lose weight. This is a plan developed by a major food manufacturer, so there's plenty of money behind it and you can design the experiment

with whatever resources you think are needed. For example, you can have a large number of participants and a control group; you can have random assignment; you can pay participants; and so on. Write up the design and discuss how you plan to address various threats to the validity of the experiment, including regression to the mean, the John Henry effect, maturation, history, and so on.

3. Look through your local newspaper and find a report of some innovative social program, such as prison reform, a program to provide jobs to young people during school vacations, or a proposal for a new welfare system. Describe how you would evaluate the success of the innovative program. First, briefly describe the innovative program. Next, specify the key dependent variable that the program seeks to affect. Finally, describe how you would measure the dependent variable and how you would assess whether the program had any effect on that variable.

4. A local government agency has awarded you a contract to evaluate a program it has had in place for two years. The program is a day-care center for children of working mothers in a predominantly Spanish-speaking area of a major city in the Southwest. When the program was proposed, its advocates claimed that the cost, in dollars, would be less than the cost of welfare payments to the women who couldn't work because they had no place to leave their children. Now, two years later, the data are in and the bottom line is that it costs much more to keep each child than it would cost to close the day-care center and return to the system of direct welfare payments to the mothers.

   Advocates for the center claim that there are many more concerns than just "the bottom line," but they are having a hard time articulating those concerns. Design an evaluation research effort that addresses all the concerns, *except the fiscal ones*, of the parties in this situation.

5. Assume that you have developed a study technique that you believe will result in students scoring higher on this exam. You test the technique with the following design:

   R $O_1 \times O_2$

   R O_3 O_4

   List *all* the predictions you can make. If it turns out that O_4 is greater than O_1, what will you conclude? Describe the confounds to internal validity in this experiment. Be sure to distinguish between internal and external validity.

Further Reading

Research design. Brink and Wood (1998), T. D. Cook and Campbell (1979), Glass et al. (1979), Kazdin (1998), Kirk (1995), Marczyk et al. (2005), Trochim (1986). On design in qualitative research: Creswell (2003), Lincoln and Guba (1985), Maxwell (2005).

Case study methods. Feagin et al. (1991), Gerring (2007), Gomm et al. (2000), Mills et al. (2009), Yin (2003).

Thought experiments. Brown (2010), Horowitz and Massey (1991), Sorensen (1992), Tindale and Vollrath (1992).

Diffusion of responsibility. Bandura et al. (1996), Corrion et al. (2009), Freeman et al. (1975), Henriksen and Dayton (2006), Shotland and Straw (1976).

Naturalistic experiments. McGarva et al. (2006), Ruback and Juieng (1997), van Straaten et al. (2008), Walker (2007).

Small-world research. Barabási (2002), Bernard and Killworth (1979), Cho and Fowler (2010), Crossley (2005), Kilduff et al (2008), Kochen (1989), Schnettler (2009), Shotland (1976), Watts and Strogatz (1998, 1999, 2003, 2004).

The lost-letter technique. Ahmed (2010), Bridges et al. (2002), Stern and Faber (1997), Waugh et al. (2000).

Deception and debriefing. Barchard and Williams (2008), Benham (2008), Broeder (1998), K. S. Cook and Yamagishi (2008), Fisher (2005), Kimmel (1998), F. G. Miller et al. (2008), Nicks et al. (1997), Ortmann and Hertwig (1997, 1998), Sharpe and Faye (2009), Sieber et al. (1995) Taylor and Shepperd (1996). See also: **Further Reading** on deception in field studies, Chapter 14.