CHAPTER 7

RANDOMIZED CONTROLLED EXPERIMENTS AND CAUSAL INFERENCE

EXPERIMENTATION IN SCIENCE

In chapter 1 we said that the orientation of experimental research is the search for causes, and in this chapter we begin by describing typical experimental designs that use the principle of randomization to assign the sampling units to groups or conditions. Next, after noting several features and limitations of randomization, we turn to the philosophical puzzle of causality. As one author commented, "Though it is basic to human thought, causality is a notion shrouded in mystery, controversy, and caution, because scientists and philosophers have had difficulties defining when one event *truly causes* another" (Pearl, 2000, p. 331). We review the justification for the use of control groups in randomized experiments, and after a brief discussion of a historically significant contribution to control group design by Richard L. Solomon, we turn to the work of Donald Campbell and his associates that has focused on an assortment of threats to the validity of causal inferences. We conclude this chapter with a discussion of subject-related and experimenter-related artifacts and the strategies used to control for them.

Although the primary emphasis of this chapter is on randomized controlled experiments, we generally use the term *experimentation* in a broad sense rather than restricting it only to randomized experiments; this broad usage is also frequent in

science. Shadish, Cook, and Campbell (2002) recommended that *experimentation* be understood as referring "to a systematic study design to examine the consequences of deliberately varying a potential causal agent" (p. xvii). Four features that Shadish et al. thought representative of all experiments are "(1) variation in the treatment, (2) posttreatment measures of outcomes, (3) at least one unit on which observation is made, and (4) a mechanism for inferring what the outcome would have been without treatment" (p. xvii). As an illustration of this broad conception, students who have taken a course in chemistry know that the typical experiment consists of mixing reagents in test tubes. How does this process fit in with Shadish et al.'s four features of experiments? The variation in treatment is analogous to an implicit comparison with other combinations of reagents; the posttreatment measure is the resulting compound; the observations typically require calibrated measurements; and the outcome without treatment is analogous to the status of the reagents before they were combined. As another illustration of nonrandomized experiments, an article in *Physics World* invited readers to nominate the "most beautiful experiments of all time" (G. Johnson, 2002). None of the top ten involved randomization. The experiment ranked first was not even an empirical demonstration, but a classic thought experiment (by Thomas Young) to demonstrate that light consists of particles that act like waves. The second-ranked experiment was Galileo's revelation, in the late 1500s, that because all bodies necessarily fall with the same velocity in the same medium, it follows that objects of different weights dropped from the Leaning Tower of Pisa would land at the same time.

It is not hard to find many fascinating examples of nonrandomized experiments in psychology as well. For instance, Jose M. R. Delgado (1963) showed that the electrical stimulation of various brain regions resulted in decreased aggressive behavior. In one study, the boss monkey (named Ali) in a colony of monkeys that lived together had an electrode inserted in his caudate nucleus. The switch that turned on the current to Ali's electrode (via a tiny radio transmitter) was available to the other monkeys. They learned to approach and press the switch whenever Ali began to get nasty, causing him to become less aggressive immediately. In a more dramatic demonstration, Delgado got into a ring with a fierce bull whose brain had been implanted with electrodes. Just as the bull was charging toward him, Delgado turned on radio-controlled brain stimulation, causing the bull to stop in midcharge and become passive. Few experimenters have opportunities to demonstrate such confidence in their causal generalizations. In the next chapter we will have more to say about experiments that focus on only one unit (called a *case*), or on only a few units, and in which randomization is seldom used (called *single-case experiments*).

RANDOMIZED EXPERIMENTAL DESIGNS

By far the most common randomized controlled experiments are those in which the sampling units are assigned to receive one condition each, called a **between-subjects design** (or a **nested design,** as the units are "nested" within their own groups or conditions). In biomedical research, these experiments are regarded as the "gold standard," such as randomly assigning the subjects to receive an experimental drug or a **placebo** (a substance without any pharmacological benefit that is given as a pseudomedicine to subjects in a control group). This particular design is shown in Table 7.1, in which the

TABLE 7.1
Simplest between-subjects design

New drug	Placebo
Subject 1	Subject 2
Subject 3	Subject 4
Subject 5	Subject 6
Subject 7	Subject 8
Subject 9	Subject 10

randomly assigned subjects of the experimental group are arbitrarily labeled 1, 3, 5, 7, 9, and the randomly assigned subjects of the control group are arbitrarily labeled 2, 4, 6, 8, 10. Suppose, however, that the drug in question is intended to treat a terminal illness. Assigning some of the subjects to receive only a placebo would deny them access to a potentially life-extending treatment. According to an international declaration adopted in Helsinki in October 2000 by the general assembly of the World Medical Association, placebos may be used only when there are no other drugs or therapies available for comparison with a test procedure. Thus, one option is to give control subjects not a placebo, but the best currently available treatment, so that the comparison is between the experimental drug and the best available option. Another declaration, issued in 2002 by the Council of International Organizations of Medical Sciences (CIOMS), stated that the only ethical exception to *not* using an effective treatment as the control condition is if there is no such treatment available and it is unlikely to become available in the foreseeable future (Levine, Carpenter, & Appelbaum, 2003).

Our earlier discussion of the ethics of placebo control groups (chapter 3) noted that another possibility in some situations is to use a **wait-list control group**—assuming, of course, that the design has been approved by an ethical review committee. For example, there might be an alternative treatment, that is unavailable for logistic or cost reasons, thus leaving the placebo condition as our *only* viable option. The simplest randomized wait-list design using a placebo control consists of two conditions:

Group 1	R	O	X	O		O
Group 2	R	O		O	X	O

where R denotes random assignment, O denotes observation (i.e., measurement or testing), and X denotes the experimental treatment. In this example, the participants randomly assigned to Group 1 are given the drug (X) over a specified period, during which the participants assigned to Group 2 (the wait-list control group) receive not the drug or any alternative medicine, but a placebo. Once it becomes clear that the new drug is effective, the trial is terminated, and the wait-list group is given the new drug.

If there are a sufficient number of participants in the wait-list control condition, and if it would not violate the international declarations or be hazardous to the subjects'

health to wait to receive the new beneficial drug, another possibility is to introduce different periods of delay before the new drug is administered to these subjects:

Group 1	R	O_1	X	O_2		O_3		O_4		O_5
Group 2	R	O_1		O_2	X	O_3		O_4		O_5
Group 3	R	O_1		O_2		O_3	X	O_4		O_5
Group 4	R	O_1		O_2		O_3		O_4	X	O_5

Groups 2, 3, and 4 are subgroups of the wait-list control condition. Once it is clear that the experimental drug (X) is effective, Groups 2, 3, and 4 receive the drug at different intervals. Repeated measurements (O) allow us to gain further information about the temporal effects of the drug by comparing O_2, O_3, O_4, and O_5 in Group 1 with the same set of observations in the wait-listed controls.

In between-subjects studies in which the participants are asked to make subjective judgments, their implicit ranges may be different if they are not operating from the same emotional, psychological, or experiential adaptation levels or baselines (Helson, 1959; cf. Birnbaum, 1999). Consequently, we might wish to have the subjects make repeated judgments based on different conditions, called a **within-subjects design** because each participant receives more than one condition (also called a **crossed design** because the subjects are thought of as "crossed" by conditions). A problem with within-subjects designs, however, is that the order in which the conditions are administered to the subjects may be confounded with the condition effect. Suppose there are four conditions (A, B, C, and D) to be administered to young children who are to be measured following each condition. The children may be nervous when first measured, and they may respond poorly. Later, they may be less nervous, and they may respond better. To address the problem of systematic differences between successive conditions, the experimenter can use **counterbalancing,** that is, rotating the sequences of the conditions (as illustrated in Table 7.2) in what is called a **Latin square.** Notice that it is a square array of letters (representing the conditions) in which each letter appears once and only once in each row and in each column. In this illustration, all four

TABLE 7.2
Latin square design

	Order of administration			
	1	2	3	4
Sequence 1	A	B	C	D
Sequence 2	B	C	D	A
Sequence 3	C	D	A	B
Sequence 4	D	A	B	C

TABLE 7.3
2 × 2 Factorial design

A1. Experimental treatment		A2. Placebo control	
B1. Women	B2. Men	B1. Women	B2. Men
Subject 1	Subject 11	Subject 2	Subject 12
Subject 3	Subject 13	Subject 4	Subject 14
Subject 5	Subject 15	Subject 6	Subject 16
Subject 7	Subject 17	Subject 8	Subject 18
Subject 9	Subject 19	Subject 10	Subject 20

conditions are administered to the children in a counterbalanced pattern, so that the children randomly assigned to Sequence 1 receive conditions in the sequence A, then B, then C, and finally D. In Sequences 2 through 4, the conditions are administered in different sequences, BCDA, CDAB, and DABC, respectively.

Another popular arrangement of conditions in experimental research is called a **factorial design** because there is more than one variable (or factor) and more than one level of each factor. Suppose that women and men are randomly assigned to a drug or a placebo group. We have a two-factor design with two levels of the factor of gender (women vs. men) and two levels of the manipulated variable (drug vs. placebo), with the assignment of the subjects illustrated in Table 7.3. This arrangement is described as a 2 × 2 factorial design (2 × 2 is read as "two by two") or 2^2 factorial design. Of course, factorial designs are not limited to only two factors or to only two levels of each factor. We will have much more to say about the analysis of factorial designs in chapters 16–18, but one common procedure in the case of the design shown in Table 7.3 is to compute a 2 × 2 analysis of variance in which we analyze the between-group variation of the drug versus the placebo condition (A1 vs. A2), the between-group variation of women versus men (B1 vs. B2), and the interaction of factors A and B. However, if we are primarily interested in some predicted pattern of all four condition means, we might prefer to address this prediction by means of a 1 × 4 contrast, as described in chapter 15.

Other variations include **fractional factorial designs,** also called *fractional replications* (Winer, et al., 1991), which use only some combinations of factor levels (rather than using all combinations of all factor levels, also known as **full factorials**), and **mixed factorial designs,** consisting of both between- *and* within-subjects factors. An experiment in which women and men (a between-subjects factor) both received a sequence of treatments (and were measured after each one, so that "treatments" is now operationalized as a within-subjects factor) is an example of a mixed factorial design.

CHARACTERISTICS OF RANDOMIZATION

In their classic statistics text, George W. Snedecor and William G. Cochran (1989) made the point that "randomization gives each treatment an equal chance of being allotted to any subject that happens to give an unusually good or unusually poor

response, exactly as assumed in the theory of probability on which the statistical analysis, tests of significance, and confidence intervals are based" (p. 95). Earlier, in another classic text, R. A. Fisher (1971) had noted that by the use of "the full procedure of randomisation . . . the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated" (p. 19). That is, randomization is a way of dealing with unsuspected sources of bias, though, of course, it cannot guarantee that all important natural differences among the subjects will be *exactly* balanced out. Suppose we were working with pairs of subjects, one of whom in each pair is to be randomly assigned to the experimental condition by the flip of a coin. Snedecor and Cochran (1989, p. 95) noted that, with n pairs, the probability that any one treatment will be assigned to the superior member in every one of the n pairs was $1/2^{n-1}$. Thus, with 5 pairs, the probability is .06, with 10 pairs it is .002, and with 15 pairs it is .00006. In other words, as sample sizes increase, it becomes less and less likely that the treatment condition subjects will be very different from the control condition subjects even before the subjects receive the treatment condition. Those relatively rare instances in which very large differences between conditions existed even before the treatments were administered are sometimes referred to as *failures of randomization.*

Instead of flipping a coin, most researchers prefer using a computer, a scientific calculator, or a table of random digits to generate the random numbers they need. To use Table B.9 (in Appendix B), we start by blindly choosing a place in the table and then simply reading across the row or down the column. We must decide in advance which numbers will determine the units to be assigned to different conditions. Suppose we decide that odd numbers will determine the units to be assigned to the experimental group and that even numbers (including zero) will determine those to be assigned to the control group. Say we want to assign 20 subjects, and the 20 digits that we randomly select are 10097, 32533, 76520, and 13586. We number our subjects from 1 to 20 and align all subjects with their random digits, for example:

Random digit	1	0	0	9	7	3	2	5	3	3	7	6	5	2	0	1	3	5	8	6
Subject number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Then, because we want to assign 10 subjects to the treatment and 10 to the control, Subjects 1, 4, 5, 6, 8, 9, 10, 11, 13, and 16 are assigned to the treatment condition (their associated random digits are odd), and the remaining subjects are assigned to the control condition. If there are fewer than 10 odd numbers in our series of 20 random digits, we simply select the first 10 even-numbered subjects to be our control sample, and the remaining subjects become the treatment sample. Another procedure for the random assignment of subjects, or other sampling units, is simply to number the units from 1 to 20, write the numbers on slips of paper, put them in a bag, shake the bag well, and then blindly draw a slip for each participant.

In chapter 15 we will discuss the allocation of different numbers of units to the samples, usually because of cost factors and because only one or two specific patterns of effects are of primary interest. This approach is traditionally known as *optimal design* in statistics, and the goal of the differential allocation of sample sizes is to increase statistical precision (Kuhfield, Tobias, & Garratt, 1994; McClelland, 1997).

Although we may sometimes want to use this procedure, it may also limit our options to discover something new and interesting from a "nonoptimal" design. We will illustrate all these ideas in the later chapter.

Although randomization is also intended to avoid the problem that experimenters might, even if unconsciously, let their feelings influence their decisions about which subjects will be placed in different conditions (Gigerenzer, Swijtink, Porter, Daston, Beatty, & Krüger, 1989), perfect experimental control is an ideal that can be elusive. Suppose the problem is not that experimenters' feelings are suspected of having gotten in the way, but that the sample sizes are small and it simply happens that the experimental and control groups are not comparable to begin with. Say that the mean pretest scores in the experimental and control groups are 50 and 40, respectively, and the mean posttest scores in these same groups are 52 and 48. Concentrating our attention only on the posttest scores, we would conclude that the experimental group showed a positive effect. However, when we take into account the mean baseline differences, we see that the positive effect was an illusion; the mean pre-to-post gain in the experimental group was one quarter the size of that in the control group. Thus, another variation on the randomized designs noted before is to use pretest measurements to establish baseline scores for all subjects. For example, in a drug trial of a pharmaceutical that treats high blood pressure, the researchers would want to know what the subjects' blood pressures were before they received the drug. For a detailed discussion of pretest-posttest designs (particularly those used in pharmaceutical trials), including many of the advantages and limitations of the common statistical procedures that are used to analyze these designs, see Bonate (2000). Incidentally, in randomized drug trials, one concern is that the participants, fearing that they have been placed in the placebo group, may surreptitiously arrange to split and then share with other participants the pills they've received; the result may be an underestimation of the true effectiveness of the pharmaceutical treatment. Another issue is that potential subjects may be reluctant to volunteer in the first place because they are concerned about being assigned to the group that will not receive the experimental drug; their reluctance reduces the pool of available subjects and thus could seriously jeopardize the generalizability of the observed results (Kramer & Shapiro, 1984).

Before leaving this section we should also mention that, although we have used the term *random* we have not actually defined it. Dictionaries frequently propose "aimlessness" or the "lack of purpose" as synonyms of *randomness,* or they note that the "equal probability of being chosen" is a statistical characteristic of *randomness.* We will have more to say about probability and randomness in chapter 9, but it is important not to confuse randomness with aimlessness, or "hit-or-miss" sampling, which can seldom be accurately described as random in the scientific sense. We can illustrate by asking someone to write down "at random" several hundred one-digit numbers from 0 to 9. When we tabulate the 0s, 1s, 2s, and so on, we will see obvious sequences and that some numbers clearly occur more often than others; these patterns would not occur if each digit had an equal probability (10%) of being listed (Wallis & Roberts, 1956). Interestingly, psychologist Allen Neuringer (1992), using a single-case experimental strategy, was able to reinforce the behavior of pigeons to make left-right choices that looked pretty random, and he then used feedback to reinforce some individual Reed College students to generate sequences of numbers that also looked like random

sequences (Neuringer, 1996; Neuringer & Voss, 1993). As Kolata (1986) mentioned in *Science* magazine, randomness is a difficult concept even for statisticians and philosophers to define, one reason that "books on probability do not even attempt to define it" (p. 231). The problem, as one leading scholar, Persi Diaconis, observed, is that "if you look hard, things aren't as random as everyone assumes" (Kolata, p. 1069). However, Diaconis noted that some outcomes at least turn out to be "almost random," such as flipping an ordinary coin millions of times to discover any bias.

THE PHILOSOPHICAL PUZZLE OF CAUSALITY

Turning the key in a car's ignition gets the motor going; taking an aspirin when you are running a fever will usually lower your temperature; studying for an exam is better than not studying for it; pulling a sleeping dog's tail usually evokes an aggressive response; smoking cigarettes over a long period of time can cause emphysema, lung cancer, and heart disease; aesthetically applying paint to a canvas produces a painting. All these are examples of **causal relations.** That is to say, they imply a relation between a *cause,* which may be a responsible human or physical agent or force, and an *effect,* which may be an event, a state, or an object. They also emphasize what Aristotle called "efficient causation" (more about this below), which developmental psychologists now believe is perceptible even in infants (Schlottmann, 2001). Interestingly, some historians and others have argued that, for thousands of years, well before the Golden Age of Greece, the idea of effects caused by objects or physical forces was hardly of much interest. The reason, they say, is that events and personal ordeals were presumed to be either under the control of a divine will (physical catastrophes were expressions of angry gods, and felicitous events were expressions of benevolent gods) or attributed to human or animal agents (e.g., a hunter or a forager gathering food, or an animal trampling a person to death).

By the fifth century B.C., Greek philosophers (such as Parmenides, Anaxagoras, and Empedocles) had begun to explore the problem of causality more deeply, particularly with reference to the elemental origin of the universe and the nature of change, on the assumption that change requires a cause of some kind. A century later, a major conceptual advance occurred when Aristotle differentiated four types of causation: material, formal, efficient, and final (Wheelwright, 1951). The *material cause* refers to the elementary composition of things. The *formal cause* is the outline, conception, or vision of the perfected thing. The *efficient cause* is the agent or moving force that brings about the change. The *final cause* (or *teleological explanation*) refers to the purpose, goal, or ultimate function of the completed thing. For Aristotle, the ultimate teleological explanation for everything that undergoes change was divine will and, thus, God's final plan. To recast the four causes in a modern context, we might ask, "What causes a skyscraper to be built?" The material cause would be the concrete, bricks, and steel; the formal cause, the architect's blueprint; the efficient cause, the physical work of the architect, the contractor, and the laborers and their tools; and the final cause (i.e., the ultimate objective), a building for people to occupy.

By the 13th and 14th centuries A.D., we find a symbiosis of the emerging idea of causal inference and the empirical method of "experimentation," that is, in its broadest interpretation (Rosnow, 1981). It is during this period that we see a further emphasis on efficient causation. For example, Robert Grosseteste, at Oxford University, proposed that the forces or agents (the efficient causes) accountable for the nature of light are subject to manipulability and critical observation, thereby establishing optics as both an empirical and an experimental science (Wallace, 1972). Grosseteste's visionary idea was subsequently embraced by his student, Roger Bacon, who stressed the "very necessity" of empirical experimentation and the essential role of quantification. In his *Opus Maius,* published in 1267, Bacon stated that "if . . . we are to arrive at certainty without doubt and at truth without error, we must set foundations of knowledge on mathematics, insofar as disposed through it we can attain to certainty in the other sciences, and to truth through the exclusion of error" (Sambursky, 1975, p. 155). Of two ways of acquiring knowledge of efficient causation, by reason and by experimentation (or *experience* in Bacon's words), he argued that it was only by experimentation that we can demonstrate what reason teaches us. For example, a person who has never seen fire may be persuaded by a well-reasoned argument that fire burns and injures things and destroys them, but until he has witnessed an experimental demonstration of combustion, he cannot accept what he learned as an indisputable fact, for "reasoning does not suffice, but experience does" (W. S. Fowler, 1962, p. 36).

Although medieval science seemed to explain everything, it was nonetheless a tangle of unwieldy, interlaced, cumbersome theories. Revolutionary developments occurred from the 16th to the 17th century that changed all this by undermining ancient intellectual foundations and replacing them with a theoretical foundation of mechanics that explained the action of forces on bodies. In his biography of Isaac Newton, James Gleick (2003) likened the scientific revolution that began in the 16th century with Nicolaus Copernicus, the Polish astronomer, and "staggered under the assaults of Galileo and Descartes and finally expired in 1687, when Newton published a book" (p. 49) to an "epidemic spreading across the continent of Europe during two centuries" (p. 48). Galileo Galilei, by virtue not only of his empirical methodology but also of his mathematical formulations of causal experimental effects, was a seminal figure in the development of the experimental method of science. In 1609, he made his first observations with the use of a telescope, constructed by "inserting spectacle makers' lenses into a hollow tube" (Gleick, p. 49), providing powerful evidence to support Copernicus's heliocentric theory of the universe (which in the 17th century was still at odds with ecclesiastical doctrine on the incorruptibility of the heavens). Galileo's mathematical definitions of velocity and acceleration and their dependence on time, his kinematic laws, and his laws concerning the oscillation of the pendulum—all these and other insights inspired the development of an experimental science of mechanics, emphasizing efficient causation. In particular, Galileo's *Dialogo* of 1632 and his *Discorsi* of 1638 were a reservoir of ideas that were tested experimentally by himself and others.

Nearly coinciding with Galileo's work were the contributions of the French philosopher René Descartes and his profoundly mechanistic view of nature, which, he argued, provided logical proof of the existence of God. In his *Principia Philosophia,* published in 1644, Descartes argued that everything that exists requires a cause, and thus, the physical world can be likened to a complex machine and its parts. "As regards the general cause," he argued, "it seems clear to me that it can be none other

than God himself. He created matter along with motion and rest in the beginning; and now, merely by his ordinary co-operation, he preserves just the quantity of motion and rest in the material world that he put there in the beginning" (Sambursky, 1975, p. 246). Thus, Descartes reasoned, in the same way that the complexity inherent in the design and working parts of a complicated machine confronts us with wonder about the nature of its human creator, the machinery that constitutes the universe demands that we consider the majesty of its divine creator, God.

Isaac Newton, born in the year of Galileo's death, extended the mechanistic conception by his own powerful theories on the mechanics of motion while passionately championing the experimental method of science. Like Galileo and Descartes, Newton believed in the uniformity of nature, an idea he developed in detail in *Philosophiae Naturalis Principia Mathematica* (cited as *Principia*), published by the Royal Society in 1687. Proceeding on the premise that "space is absolute," his three axiomatic laws of motion replaced the clutter of Aristotelian and medieval science with a compact mechanical theory that stood unchallenged for over 200 years. His love of the experimental method of science was no less intense than Galileo's, and in *Principia,* Newton wrote,

> For since the qualities of bodies are only known to us by experiments, we are to hold for universal all such as universally agree with experiments; and such as are not liable to diminution can never be quite taken away. We are certainly not to relinquish the evidence of experiments for the sake of dreams and vain fictions of our own devising; nor are we to recede from the analogy of Nature, which is wont to be simple, and always consonant to itself. (longer passage in Sambursky, 1975, p. 303)

In his second great work, *Opticks—or, a Treatise on the Reflections, Refractions, Inflexions and Colours of Light,* published in 1706, Newton described his famous *Experimentum Crucis* on light and color. He directed a sunbeam of white light through a triangular prism to produce a rainbow of colors (an old, but inexplicable, phenomenon) and then sent a single colored beam through another prism to show unequivocally that "white light is a mixture, but the colored beams are pure" (quoted in Gleick, 2003, p. 80).

CONTIGUITY, PRIORITY, AND CONSTANT CONJUNCTION

As word of the *Principia* spread, Newton's monumental formulation was embraced by philosophers who contended that the "new philosophy of mechanism" was also quite adequate to comprehend purposive human behavior (cf. Lowry, 1971; Wilkes, 1978). Inspired by bold extrapolations, arguments began to be made that almost everything in the panoply of natural and human phenomena could be understood by means of "Newtonianism" and the laws of motion. Humans, after all, are merely the product of biological and social engineering, a complex piece of machinery that efficiently modifies force or energy (cf. Aleksander, 1971). However, it was David Hume, the great 18th-century Scottish philosopher, who, in the words of Judea Pearl (2000), "shook up causation so thoroughly that it has not recovered to this day," for, according to Hume, the sensation of causation was "almost as fictional as optical illusions and as transitory as Pavlov's conditioning" (p. 336). Hume reasoned that, because it is intrinsic in human nature to expect the future to be like the past, the mind is essentially programmed to perceive causal links even though they are mere illusions "deriv'd from nothing but custom" (Hume, 1978, p. 183). Motion is lawful, but causality is in the mind's eye, conditioned by sensory repetitions and the mechanics of the association of ideas.

In his *Treatise of Human Nature* (subtitled *An Attempt to Introduce the Experimental Method of Reasoning Into Moral Subjects*), Hume listed a set of "rules" for defining causality. Eight years later, in his *Inquiry Concerning Human Understanding,* he argued that the appropriate method of adducing causality was Newton's, because it is by the experimental method that we "discover, at least in some degree, the secret springs and principles by which the human mind is actuated in its operation" (Hume, 1955, p. 24). Hume's eight "rules by which to judge of causes and effects" (Hume, 1978, pp. 173–175) could be boiled down to three essentials: First, "the cause and effect must be contiguous in space and time" (called *contiguity*). Second, "the cause must be prior to the effect" (called *priority*). Third, "there must be a constant union betwixt the cause and effect" so that "the same causes always produce the same effect, and the same effect never arises but from the same cause" (called a *constant conjunction*).

To illustrate, Hume (1978, p. 148) gave the example of a man who is hung from a high tower in a cage of iron. Even though he is aware that he is secure from falling, he nevertheless trembles from fear. The reason he trembles has to do with his "custom" of associating contiguous events in a causal sequence (the ideas of "fall and descent" and "harm and death")—which modern psychologists would call an *illusory correlation* of events (Fiedler, 2000). Another favorite example of Hume's was of a billiard ball lying on a table with another ball rapidly moving toward it; they strike, and the ball that was previously at rest is set in motion. This, he stated, "is as perfect an instance of the relation of cause and effect as any which we know, either by sensation or reflection" (Hume, 1978, p. 649). And yet, all that can be confidently said, he argued, is that

> the two balls touched one another before the motion was communicated, and that there was no interval betwixt the shock and the motion. . . . Beyond these three circumstances of contiguity, priority, and constant conjunction, I can discover nothing in this cause. . . . In whatever shape I turn this matter, and however I examine it, I can find nothing farther. (pp. 649–650).

Of course, merely because a physically or temporally contiguous event invariably precedes another event—and thus predicts the event well—does not automatically implicate one as the cause of the other. The barometer falls before it rains, but a falling barometer does not *cause* the rain (Pearl, 2000, p. 42). Another example (discussed by Edmonds & Eidinow, 2001) that consumed Cambridge University philosophers in the 1940s involved the idea of two factories in two towns in England, one in the south and the other in the north, but both in the same time zone. Each factory has a hooter that signals the end of the morning shift at exactly twelve noon, and every time the northern hooter sounds, the southern workers lay down their tools and exit. Although we see, as Hume might have said, *contiguity, priority,* and a *constant conjunction* of events, the northern hooter is obviously not the *cause* of the southern workers' stopping

work. It appears that what baffled the Cambridge philosophers was how to clarify, unequivocally, the essential difference between *coincidentally* and *causally* linked events. Monday precedes Tuesday, just as night precedes day; these are not *coincidental* linkages, but it would be equally absurd to say that Monday *causes* Tuesday or that night *causes* day. Some psychologists say the missing ingredient that explains the perception of causality is simply the ability to connect events by some plausible mechanism, whereas others argue that mere prediction is the necessary mechanism. Neither is a particularly satisfying answer, however. As Schlottmann (2001) observed, "Even infants have perceptual sensitivity to the causal structure of the world" and "if we always relied on mechanisms we would be locked into prejudice, having no way to go beyond what we already know" (pp. 111, 115). Edmonds and Eidinow (2001) concluded by asking: What is causality, then, merely "a furtive, cloak-and-dagger agent, never seen or touched?" or is it "a chimera, a trick played on us by our imagination" (p. 65)?

FOUR TYPES OF EXPERIMENTAL CONTROL

To make this long story a little shorter, we skip to the 19th century and British philosopher John Stuart Mill. Like Hume, Mill was skeptical about ever removing all possibility of doubt when speaking of causality. However, he reasoned that demonstrating empirically that there are both necessary *and* sufficient conditions of a presumed causal relation would produce by far the most convincing evidence of efficient causation. Mill's methods for demonstrating these necessary and sufficient conditions became the basis of an important empirical strategy of causal explanation in science, the use of *control conditions.* Before we explore this particular application, we pause briefly to note three other uses of the term *control* in experimental research. In this discussion we draw on the writing of Edwin G. Boring (1954, 1969), who was well known for his classic texts on the history of experimental psychology (Boring, 1942, 1957).

Boring (1954) noted that the original meaning of *control* was "check," because the word *control* was the diminutive of *counterroll (contre-rolle),* the term for a "duplicate register or account made to verify an official or first-made account and thus a check by a later roll upon the earlier" (p. 573). Apparently, the idea of a check (or test observation) to verify something first came into scientific parlance during the last half of the 19th century, and by 1893, *control* also was used to refer to "a standard of comparison" (p. 574). A variation on the idea of a check or restraint is implicit in the idea of control referring to the "constancy of conditions" in an experimental research situation. For example, unless a scientist wants to study the effect of extreme temperature variation, it would not be advisable to allow the temperature in a laboratory to vary capriciously from very chilly to very hot. If such variation occurs, the scientist will be unable to claim the constancy of conditions that allows certain statements of cause-and-effect relations to be made. To avoid this problem, the scientist *controls* (i.e., holds in check) the laboratory temperature by keeping it constant, removing the possibility of systematic error variability leading to spurious correlations and errant conclusions.

Two other contemporary uses of the term *control* can be found in psychophysical research and in single-case research on behavioral learning principles. In

psychophysical research, experimenters use the term *control series.* To illustrate, suppose that blindfolded subjects are asked to judge whether their skin is being touched by one or two fine compass points. When two points are very close to one another, they should be perceived as only one point. A control series of psychophysical trials would consist of varying the distance between the two points and presenting one point on a certain percentage of trials. In this way, it is possible to identify the smallest detectable distance between two points while controlling for suggestibility. That is, if the subjects believe they are *always* being touched by two points, they might never report being stimulated by only one point. In single-case experimental research (discussed in the next chapter), the term *behavior control* is often used; it refers to the shaping of learned behavior based on a particular schedule of reinforcement designed to elicit the behavior in question.

As Boring (1954) noted, although the use of the term *control* to refer to "a standard of comparison" is of relatively recent origin in the history of science, the idea can be deduced from John Stuart Mill's work and is also implicit in much earlier work (e.g., F. P. Jones, 1964; Ramul, 1963). For example, F. P. Jones (1964) mentioned an example going back to the Greek philosopher Athenaeus in the second century A.D., in which he described how a magistrate in ancient Egypt had discovered citron to be an antidote for poison. According to the story, the magistrate had sentenced a group of criminals to be executed by being exposed to poisonous snakes. Although the sentence was carried out with due diligence, it was reported back to him that none of the prisoners had died. What apparently had happened was that, while they were on their way to the place where they were to be executed, a market woman took pity on them and gave them some citron to eat. The next day, on the hypothesis that it must have been the citron that had saved their lives, the magistrate had citron fed to one of each pair of criminals and nothing to the others. Exposed to poisonous snakes a second time, those prisoners who had eaten the citron survived and those not given it died instantly. This story not only illustrates the early use of a control group but also provides another example of serendipity as well as the early use of replication, for Athenaeus noted that the experiment was repeated many times to firmly establish that citron was indeed an antidote for poison.

MILL'S METHODS OF AGREEMENT AND DIFFERENCE

John Stuart Mill proposed four methods of experimental inquiry in his 1843 classic, *A System of Logic, Ratiocinative and Inductive,* but it was his methods of agreement and difference that best provide the logical basis of the use of a control group or comparison condition in simple randomized controlled experiments. The **method of agreement** states, "If X, then Y," X symbolizing the presumed cause and Y the presumed effect. The statement means that if we find two or more instances in which Y occurs, and if only X is present on each occasion, X is implied as a **sufficient condition** of Y. In other words, X is *adequate* (i.e., capable or competent enough) to bring about the effect (Y). In baseball, for example, we would say there are several sufficient conditions for scoring runs, such as hitting a home run (X_1), stealing home (X_2), a wild pitch with the bases loaded (X_3), a hit that moves a runner home (X_4), and so

forth. Another example would be sufficient conditions for starting a fire (Y), which include striking a match (X_1), using a flint to create a spark that ignites dry leaves (X_2), turning on a gas stove (X_3), or just waiting in a storm for lightning to strike (X_4). In social psychology, an example of sufficient conditions might be those that are adequate to cause a person to pass a rumor, for example, seeking confirmation of information (X_1), manipulating a situation by sending up a trial balloon (X_2), impressing someone with one's privileged position (X_3), trying to convince people to conform to a set of group or societal norms (X_4), or trying to manipulate stock prices (X_5).

The **method of difference** states, "If not-X, then not-Y." It means that, if the presumed effect (Y) does not occur when X is absent, then we suspect X is not just a sufficient condition of Y, but a **necessary condition** for Y to occur In other words, X is *indispensable,* or absolutely essential in order for Y to occur. For example, to win in baseball (Y), it is *necessary* to score more runs (X) than the other team. Though we noted that there were several (sufficient) ways of scoring runs, the fact is that not scoring any runs (not-X) will inevitably result in not winning (not-Y). Similarly, we all know that oxygen is a necessary condition of fire, for without oxygen (not-X) the fire goes out (not-Y). The necessary condition for rumor generation appears to involve an optimum combination of uncertainty and anxiety, as rumors are essentially suppositions intended to make sense of ambiguous events or situations that make us feel apprehensive or nervous about what to believe and how to behave appropriately (Rosnow, 2001).

Table 7.4 illustrates the idea of necessary and sufficient conditions in still another situation. Suppose we are told that five people have been diagnosed with food poisoning. After some probing, we discover that all five people reported

TABLE 7.4
Illustration of agreement and difference methods

Persons	Ate burger	Ate tuna sandwich	Ate fries	Ate salad	Drank shake	Got food poisoning
Mimi	Yes	No	Yes	No	No	Yes
Gail	No	No	No	Yes	Yes	No
Connie	No	No	Yes	No	No	No
Jerry	No	Yes	No	Yes	No	No
Greg	No	Yes	No	No	Yes	No
Dwight	No	No	No	Yes	No	No
Nancy	Yes	No	Yes	Yes	No	Yes
Richard	No	Yes	Yes	Yes	No	No
Kerry	No	No	No	Yes	No	No
Michele	Yes	No	Yes	Yes	Yes	Yes
John	Yes	No	Yes	Yes	No	Yes
Sheila	Yes	No	No	No	No	Yes

Note: Based on a similar example in *Logic and Philosophy: A Modern Introduction* (6th ed.), by H. Kahane, 1989, Belmont, CA: Wadsworth. Used by permission of Howard Kahane and Wadsworth Publishing Co.

having eaten in a fast-food restaurant; seven others, who also ate at the same place around the same time, did not get sick. This table lists the foods consumed by each of them. Notice that the burger appears in each case of food poisoning, but no other food was consistently present in each case of food poisoning. This evidence suggests that the burger was the necessary and sufficient causal agent. However, the manager tells us that one of the waiters left early on the day in question, after complaining of dizziness and nausea. Thus, we have an alternative hypothesis to the burger hypothesis, which is that the waiter was the necessary condition and the foods he touched were the sufficient conditions. Going through the checks given to customers that day, the manager thinks that some of those identified as having food poisoning were probably not served by this waiter, and others who did not become ill *were* probably served by him. If correct, this assessment rules out the waiter hypothesis and leaves us with only the burger hypothesis. Because the burger was present (X) in every case in which food poisoning (Y) occurred, and the burger was absent (not-X) in every case in which food poisoning did not occur (not-Y), we conclude that the burger was necessary and sufficient to produce the outbreak of food poisoning.

BETWEEN-GROUP DESIGNS AND MILL'S JOINT METHOD

To understand the application of Mill's two methods to the logic of causal inference in a randomized controlled experiment, imagine that X represents a tranquilizer that can be obtained without prescription, and Y represents a reduction in measured tension. Say we give a group of people who complain of tension a certain dosage of the drug, and they show a reduction in measured tension. Can we now conclude from this observation that it was the tranquilizer that led to the reduction in tension? Not yet, because what we require is a control condition (a not-X condition) with which to compare the reaction in the drug group. In other words, we need a group of similar subjects to whom we do not give drug X. On the assumption that these subjects are, in fact, similar to those in the drug group in all respects except for the absence of X, then finding no reduction of tension (i.e., not-Y) in the control condition would lead us to conclude that taking drug X is an effective tension reducer and that not taking it (or an equivalent drug) will result in no observable reduction in tension.

Notice that the group given the drug, the *experimental condition,* resembles Mill's "If X, then Y" method of agreement, whereas the group not given the drug (the *control condition*) resembles his "If not-X, then not-Y" method of difference. When viewed together in this way, Mill's two methods are collectively referred to as the **joint method of agreement and difference.** Mill believed the joint method could be generalized to many different situations in which we use empirical observation and reason to rule out some hypotheses and argue for others. He realized, however, that other logical stipulations may be required to make the most solid case for causation. In the example we have been discussing, although we are on safer grounds to conclude that taking the drug (X) is what led to tension reduction (Y), it is necessary to stipulate that "taking the drug" means something different from just getting a chemical into people's bloodstreams. "Taking the drug" means, among other things, (a) having someone give people a pill,

(b) having someone give them the attention that goes with pill giving, (c) having them believe that a relevant medication has been administered, and (d) having the ingredients of the drug find their way to the subjects' bloodstreams.

Usually, when testing a new medication, the researcher is interested only in the subjects' physical reactions to the active ingredients of the medication. The researcher does not care to learn that the subjects will get to feel better if they *believe* they are being helped, because this fact (i.e., the power of suggestion) is already established. But if the researcher knows this, then how is she or he to separate the effects of the chemical ingredients of the drug from the effects of pill giving, subjects' expectations of help, and other psychological variables that may also be sufficient conditions of *Y*? The answer is by the choice of a different (or an additional) control condition. For example, in Table 7.1 we showed a between-groups design in which the control group was given a placebo rather than given nothing. If there were an appropriate drug already on the market, then the Helsinki declaration (mentioned earlier in this chapter) stipulates that the controls must receive the appropriate alternative treatment. Nonetheless, it has long been a convention in drug research that biomedical experimenters routinely use placebo control conditions. The general finding, incidentally, is that placebos are often effective, and sometimes even as effective as the far more expensive drugs for which they serve as the controls. Of course, in order to tease out the effect of the placebo, we would also need to control for the power of suggestion implicit in being given a placebo and thinking it is an active pharmaceutical. For example, when an individual has received the therapeutic drug in the past, the person may be conditioned to make "drug-anticipatory responses" to the placebo (Ramsay & Woods, 2001, p. 785). Thus, a model randomized drug experiment might have more than one control group (Dennenberg, 2002; Ross, Krugman, Lyerly, & Clyde, 1962) as well as stipulations concerning the prior treatment status of the participants.

INDEPENDENT, DEPENDENT, AND MODERATOR VARIABLES

Frequently it is clear what kind of control condition is needed, but sometimes it is not immediately apparent what kind of control group (or groups) to use. Thus, researchers rely on the wisdom of experience to decide how to frame the control condition. In the research described above, we first used a no-pill control group and then a placebo control. When there is a choice of control groups, how can the researcher decide on the most appropriate variables to control for? Two important considerations are what the particular question (or questions) of interest are and what is already known about the research area. However, even an experienced experimenter may go astray in choosing control groups when he or she makes a major shift of research areas or draws analogies that lead to faulty inferences or spurious assumptions (cf. Lieberman & Dunlap, 1979; Peek, 1977; Rosenthal, 1985a; Shapiro & Morris, 1978; Wilkins, 1984). We will have more to say shortly about teasing out causal connections in an experimental design where there is more than one control condition, but there are other considerations as well.

Hume's idea was of a "constant conjunction" between causes and effects. In actuality, there is hardly ever a perfect (or even a near-perfect) association between

the presence and absence of causes and effects. We return to this idea in chapter 11 when we give a number of examples of quite meaningful, but very small, effect size correlations in drug trials (similar to the small effect sizes in the aspirin study noted in Table 2.5 of chapter 2). Indeed, effect size correlations smaller than .10 are quite common in this area, but effect size correlations in any area may vary considerably even when similar cause-and-effect variables are observed over a period of time or in different situations. One reason for all this variability is that other factors (called **moderator variables**) may alter the relationships between cause-and-effect variables. For example, rewarding a child while she or he is in the presence of a group of other children may influence the child's attitude toward the group (A. J. Lott & B. E. Lott, 1968; B. E. Lott & A. J. Lott, 1960), but it is conceivable that variables such as the child's state of mind, personality, and degree of familiarity with the group may moderate (alter) that relationship. Similarly, it is also generally true that attributing a persuasive communication to a highly respected (as opposed to a less respected) source will improve its effectiveness in influencing attitudes, but it is also true that respectability is in the eye of the beholder. Moderator variables in this situation include idiosyncratic factors related to how each individual perceives the trustworthiness of the particular source of the communication. As another case in point, there is research that shows that the positive or negative mood people are in when subjected to persuasive communications may moderate their receptivity or resistance to those communications (Rosnow, 1968). Incidentally, it is important not to confuse moderator variables with **mediator variables;** the latter are defined as conditions, states, or other factors that intervene between the independent variable and the outcome variable in a causal chain (R. M. Baron and Kenny, 1986).

Although we have used the terms *cause* and *effect,* the terms *independent variable* and *dependent variable* are perhaps more commonly used by behavioral and social researchers. With the rise of positivism, it became unfashionable for a time to speak of causal relations. Instead, terms like *functional relations* (discussed in the next chapter) and *functional correlations* became fashionable (Gigerenzer, 1987; Wallace, 1972). Researchers might refer to the "effects of *X* on *Y*," but not to *causal effects.* By *X* and *Y,* they generally meant whatever were conceived to be the stimulus (*X*) and outcome (*Y*) conditions in a study. The term **variable** became popular because it implied that the factors of interest were subject to variation. However, as psychologist David Bakan (1967) commented, "Variables, whatever they may be *in re,* do not exist there as variables. For variables are, by definition, sets of categories; and categories are the result of someone's delineation, abstraction, and identification" (p. 54). Though the idea of a variable is further qualified by means of a distinction between the dependent variable (*Y*) and the independent variable (*X*), it should be understood that any condition or factor can be conceived of as either an independent or a dependent variable, depending on how we frame the situation or conceptualize the particular context of defined antecedents and consequences. Operationally, the **dependent variable** is the status of a measurable consequence (e.g., its presence or absence, or an increase or a decrease in a measured outcome) that presumably depends on the status of an antecedent condition, or **independent variable.**

Another reason for the use of the terms *independent variable* and *dependent variable* is that they are broad enough to encompass suspected causal agents or

conditions that are not subject to manipulation. Suppose we found a relationship between gender and height and wanted to call one an independent variable and the other a dependent variable. Common sense leads us to conclude that gender is more likely to determine height than that height is likely to determine gender, because we know that a person's gender is biologically established at conception. By changing the context, however, we can view gender as a dependent variable; for example, we can say that gender is determined by genetic factors, so the independent variable now is "genetic factors." To provide another illustration, we can also conceive of rumors as independent or dependent variables, depending on the context. For example, we know that rumors can trigger needs that instigate new rumors, which can then trigger new needs, and so on (Rosnow, 1980). Without getting drawn into murky metaphysics, suffice it to say that, as Bakan implied, all definitions of independent and dependent variables are always influenced by someone's "delineation, abstraction, and identification."

SOLOMON'S EXTENDED CONTROL GROUP DESIGN

We turn shortly to the influential ideas of Campbell and Stanley, but first, we want to mention some earlier work of experimental psychologist Richard L. Solomon. As Campbell and Stanley (1966) reminded readers, it was psychological and educational researchers between 1900 and 1920 who created the orthodox control-group design in which a pretested experimental group was compared with a control group. Designs like these were used with some frequency, usually "without need of explanation," Campbell and Stanley noted (p. 13; see also Dehue, 2000). Solomon's work represents a cutting-edge transition in thinking about control group designs in behavioral and social experimentation. In an article published in 1949, he raised the question of whether pretesting subjects in pre-post designs might have a sensitizing effect on their reactions to the experimental treatment, and he argued that orthodox two-group designs were unable to address this problem. Solomon also anticipated some ideas that were later developed in more depth by Campbell and Stanley, and though we would not endorse the specific details of all of his recommendations, we nevertheless want to recognize the historical sequence of Solomon's forward-looking work.

To put the pretest sensitization problem in context, Solomon described a popular design strategy in the field of attitude change research. The participants received a pretest that, if not identical to the posttest, was similar in terms of the scale units on the posttest. Typically there was an experimental group and a control group, and either the groups were matched on some criterion or the subjects were randomly assigned to the groups. Solomon's position was that the two-group design was deficient because of its failure to control for the possibility that merely taking the pretest could affect how the subjects responded to the experimental treatment. For example, the pretest might change their attitudinal "set" or influence some other attentional factor so that they perceived the experimental treatment differently than if they had not been pretested, and their responses to the treatment were affected accordingly. To control for this problem, Solomon cautioned researchers to use either a three-group or, preferably, a four-group design.

TABLE 7.5
Solomon's (1949) three-group design

A. Basic three-group design

Conditions	Experimental group	Control Group I	Control Group II
Pretest	Yes	Yes	No
Treatment	Yes	No	Yes
Posttest	Yes	Yes	Yes

B. Results of an experiment in spelling (n = 10)

Results	Experimental group	Control Group I	Control Group II
Pretest means	3.2	2.8	3.0 (est.)[a]
Posttest means	9.9	3.5	11.2
Improvement means	6.7	0.7	8.2

[a]Estimated from (3.2 + 2.8)/2, the average of the two available pretest means.

Note: Control Group I is easily recognized as a control for the treatment in the experimental group, but Control Group II receives the treatment and yet is called a "control group" because it controls for the presence of the pretest in the experimental group.

The three-group design that Solomon proposed appears in Part A of Table 7.5, and Part B shows the results of a spelling experiment in which 30 students in two grammar school classes were assigned to the three groups (n = 10). Pupils in the experimental group and in Control Group I were pretested on a list of words of equal difficulty (Control Group II was out of the room at the time). Then the experimental group and Control Group II were given a standard spelling lesson covering some general rules of spelling (Control Group I was out of the room), and afterward all the children were posttested on the same words as in the pretest. The (unobserved) pretest mean of Control Group II was estimated from an average of the pretest means of the experimental group and Control Group I. Solomon believed that, in order to tease out pretest sensitization, it was necessary simply to remove the combined improvement effects of Control Groups I and II from the experimental condition improvement effect, so that what remained would be either a positive or a negative effect of pretest sensitization. Solomon concentrated on the posttest means in his later work, but in this early paper (Solomon, 1949), he focused on the improvement means in Table 7.5. Then he computed $6.7 - (0.7 + 8.2) = -2.2$ and concluded that, in light of this "interaction," the "taking of the pre-test somehow diminished the effectiveness of the training in spelling" (p. 145). It was not clear why the negative effect occurred, but one possibility that he raised was that taking the pretest might have been emotionally disturbing to the children. What was clear, he concluded, was that if he had used only the experimental group and Control Group I, he would have erroneously underrated the teaching procedure, as the pretest apparently "vitiated some of the effectiveness of the teaching method" (p. 145).

TABLE 7.6
Solomon's (1949) four-group design

A. Basic four-group design

Conditions	Experimental group	Control Group I	Control Group II	Control Group III
Pretest	Yes	Yes	No	No
Treatment	Yes	No	Yes	No
Posttest	Yes	Yes	Yes	Yes

B. Numerical values needed

Results	Experimental group	Control Group I	Control Group II	Control Group III
Pretest means	A_1	A_2	A_3 (est.)[a]	A_4 (est.)[a]
Posttest means	B_1	B_2	B_3	B_4
Change means	$D_1 = B_1 - A_1$	$D_2 = B_2 - A_2$	$D_3 = B_3 - A_3$	$D_4 = B_4 - A_4$
Value of I	$D_1 - (D_2 + D_3 - D_4)$ or equivalently, $(D_1 + D_4) - (D_2 + D_3)$			

[a]Estimated from $(A_1 + A_2)/2$.

Continuing to follow the train of Solomon's thinking, Table 7.6 shows his four-group design. The additional group (Control Group III) is a control for what Campbell and Stanley subsequently called "history," or the effects of uncontrolled events that may be associated with the passage of time. Solomon (1949) mentioned certain field studies on attitude change that had been conducted during World War II; some of that research had experimented with propaganda effects. It was possible, he argued, that uncontrolled events taking place in the time between the pretest and the posttest might have impinged on all the subjects. Notice that there are two estimated pretest means (Control Groups II and III); Solomon gave them identical values based on averaging the pretest means of the experimental group and Control Group I. The value noted as D_4 is the change from the estimated pretest to the observed posttest for Control Group III that can be attributed to outside, uncontrolled events (i.e., history), he reasoned. The value of I is the difference between differences. Because there is no within-error term for the difference scores in Control Groups II and III, the ANOVA option would be to compute a 2×2 analysis on the posttest scores (with treatment vs. no treatment as one factor, and pretest vs. no pretest as the other factor), or simply to impute an error term for difference scores from error terms of the experimental group and Control Group I. Interestingly, Solomon's definition of I anticipated the most frugal way of operationalizing the 2×2 interaction, but calling the difference score of -2.2 for the results in Table 7.5 an "interaction" was simply wrong (cf. Rosnow & Rosenthal, 1989a, 1995)—we have much more to say in chapter 17 about statistical interactions in the context of analysis of variance.

Before leaving this discussion, we should note that other research has uncovered both positive and negative effects of pretesting. Solomon (1949) mentioned in a footnote that he had "preliminary evidence that the pre-test may operate to reduce post-test variance in studies of attitude change" (p. 148). In another investigation, Lessac and

Solomon (1969) used a four-group Solomon design to study the effects of sensory isolation on beagle pups. In this way they were able to estimate pretest mean scores of unpretested animals before they were placed in isolation and pretest means of their corresponding unpretested controls. Lessac and Solomon concluded that "the behavioral deficiencies found in the isolated subjects . . . must represent an active, destructive, atrophic process produced by the isolation experience" (p. 23; see Solomon & Lessac, 1968, for implications of the extended control-group design in developmental research). Using the Solomon design in an investigation of children's ability to learn state locations of large U.S. cities, Entwisle (1961), found that pretesting aided recall for the high-IQ subjects and was "mildly hindering" for the average-IQ subjects. In an attitude change study, Rosnow and Suls (1970) found that pretesting enhanced the volunteer subjects' receptivity to the experimental treatment (which involved a persuasive communication) and reduced receptivity in nonvolunteer subjects. Thus, it would appear that when a pre-post design is used in some fields (such as educational training, inducing changes in attitudes, transfer of training, performance skills) it might be prudent, as Solomon (1949) recommended, to control for the possibility of moderating effects of the pretest measurements.

THREATS TO INTERNAL VALIDITY

In chapter 4 we examined uses and definitions of the term *validity* in the context of measurement, and we now describe some additional, specialized uses of the term in the context of experimental and other research designs. In a 1963 chapter that, after the authors were inundated with hundreds of reprint requests, was published in a slightly revised version as a separate little book, *Experimental and Quasi-Experimental Designs for Research,* Campbell and Stanley introduced the terms *internal validity* and *external validity* (Campbell & Stanley, 1963, 1966). The book also stimulated considerable debate, and specific issues that Campbell and Stanley labeled one way were perceived and labeled differently by some others (Albright & Malloy, 2000). The next version of the book was greatly expanded (Cook and Campbell, 1979), in which these authors expounded on two further validity distinctions, termed *statistical conclusion validity* and *construct validity.* The most recent version of this seminal work appeared in a book coauthored by William R. Shadish, Cook, and Campbell (2002), which has continued the tradition begun by Campbell and Stanley by specifying variables and circumstances that may threaten the four types of validity not only in experimental studies but in other research as well. In this section and the two that follow, we will try to communicate a sense of this work. We should note, however, that the Shadish et al. book is encyclopedic in its coverage, reaching well beyond our abilty to summarize it in this brief discussion.

We alluded to the concept of **internal validity** in chapter 4 when we spoke of the idea of trying to rule out plausible rival hypotheses that undermine causal interpretations. That strategy is understood to be as elemental to causal inference in science as are evidence of temporal precedence and covariation. Causal inference, in other words, depends (a) not only on operationalizing a reliable relationship between an event and its presumed cause (*covariation*), as well as (b) providing some proof that the cause preceded the effect (*temporal precedence*), but also on (c) ruling out plausible rival explanations (*internal validity*). Stated still another way, the concept of internal validity is now said to imply

"the validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured" (Shadish et al., 2002, p. 38). Among several threats to internal invalidity are what Campbell and Stanley (1966) referred to as *history, maturation, instrumentation,* and *selection.* We will describe each of these in turn, but we begin with another threat to internal validity, called *regression* (a shorthand expression for "regression toward the mean"), a subject on which Campbell and Kenny (1999) wrote an entire volume.

Regression toward the mean has to do not with the actual (or raw) scores of a measure, but with the standard (or predicted) scores (Campbell & Kenny, 1999; Cohen & Cohen, 1983; Cohen, Cohen, West, & Aiken, 2003). We review standard scores (Z scores) in chapter 10, but they are raw scores from which the sample mean has been subtracted and the difference is then divided by the standard deviation. The regression equation, stated in standard score units, is $Z_y = r_{xy}Z_x$, where the standard score of Y is predicted from the product of the XY correlation (r_{xy}) times the standard score of X. Given a perfect correlation between X and Y (i.e., $r_{xy} = 1$), it follows that Z_y is equivalent to Z_x. However, if $r_{xy} < 1$, then Z_y cannot be equivalent to Z_x. For example, if $r_{xy} = .4$, then Z_y can be only 0.4 as large as Z_x. Regression toward the mean occurs when pre and post variables (X and Y) consist of the same measure taken at two points in time, and $r_{xy} < 1$. Therefore, it can be understood as a mathematical necessity whenever two variables are correlated less than perfectly. For example, finding that overweight people appear to lose weight, or that low-IQ children seem to become brighter, or that rich people appear to become poorer is a common observation in longitudinal research, but the findings might be evidence of a regression toward the mean.

Table 7.7 helps to explain how history, maturation, instrumentation, and selection affect internal validity. The table lists two of what Campbell and Stanley (1966) called "preexperimental designs" because of their relatively primitive nature. One is described in the table as a "one-shot case study" and the other, as a "one-group pre-post." One-shot case studies can be symbolized as X-O, where X denotes the exposure of a group to a variable, and O is an observation or measurement. An example of an X-O study would be introducing an educational intervention to improve reading skills and then testing the students exposed to the intervention on their reading skills. One-group pre-post studies can be symbolized as O-X-O, which means the subjects

TABLE 7.7
Four threats to internal validity in two preexperimental designs and the Solomon design of Table 7.6

	Threats to internal validity			
Design	**History**	**Maturation**	**Instrumentation**	**Selection**
One-shot case study	−	−	Not relevant	−
One-group pre-post	−	−	−	−
Solomon design	+	+	+	+

Note: A minus (−) indicates a definite weakness; a plus (+) that the source of invalidity is controlled for.

would be measured both before and after exposure to the teaching intervention. In neither preexperimental design, however, is an allowance made for a comparison with the reactions of subjects not exposed to the intervention. The minus signs in Table 7.7 imply that both preexperimental designs are totally deficient in terms of history, maturation, and selection. Only the one-group pre-post is deficient in terms of instrumentation, which is not relevant to the one-shot case study because there is no premeasurement instrument with which the postmeasurement can be compared. Also implied in Table 7.7 is that the Solomon design of Table 7.6 controls for all four threats to internal validity, as would any other randomized experiment. Now let us see how history, maturation, instrumentation, and selection are defined.

First, the term **history** implies a plausible source of error attributable to an uncontrolled event that occurs between the premeasurement (the pretest) and the postmeasurement (the posttest) and can bias the postmeasurement. History becomes a threat to internal validity when the inferred causal relationship is confounded by the irrelevant, uncontrolled event. Suppose a sudden snowstorm results in an unexpected cancellation of classes. Neither preexperimental design allows us to isolate the effects on motivation of a school closing, or to assess that variable apart from the effects of the new educational intervention designed to improve concentration. In the case of the Solomon design, there are two pretested and posttested groups, one with and the other without a treatment, so we can assess the factor of history in the treated groups apart from the untreated groups.

Second, **maturation** refers to intrinsic changes in the subjects, such as their growing older, wiser, stronger, or more experienced between the premeasurement and the postmeasurement. Maturation is a threat to internal validity when it is not the variable of interest and the causal relationship is confounded by the presence of these changes. Imagine a study in which the posttest is given 1 year after the pretest. If the students' concentration improved as a result of getting older, so they have became better at the task, neither preexperimental design could tell us whether those gains were due to students' maturing or to their being subjected to the educational innovation. The use of the Solomon design gives us an opportunity to find out how the subjects improved as a function of growing older (i.e., during the period of the experiment), as we have pre-post data on a group that did not receive the experimental treatment.

Third, **instrumentation** refers to intrinsic changes in the measuring instruments. Instrumentation is a threat to internal validity when an effect might be due to unsuspected changes in the instruments over time. In the case of the new educational innovation, we might ask whether the observed effect was due to instability (i.e., deterioration) in the achievement test or to changes in the students that were caused by the treatment. Or suppose the "instruments" were actually judges who were asked to rate the subjects. Over time, the judges might have become better raters of student concentration, so that the confounding is due not to instrument deterioration but to instrument improvement. Instrumentation bias is not a relevant issue in the one-shot case study design because the instrument is administered only once. However, it is both relevant and uncontrolled in the one-group pre-post design. It is also relevant, but specifically controlled for (i.e., identifiable), in the Solomon design, because there are two pretested groups with which we can compare the groups that were not pretested.

Finally, **selection** is a potential threat to internal validity when there are unsuspected differences between the participants in each condition. In the one-shot case study, we simply do not know beforehand anything about the state of the subjects because they are observed or measured only after the treatment has been administered. The addition of an observation before the treatment in the one-group pre-post design is a slight improvement in that it enables us to assess the prior state of the participants. The Solomon design and all other randomized experiments control for selection bias by randomly allocating participants to the groups. However, as noted earlier, random allocation is not a guarantee of comparability between groups, particularly in small-sample experiments.

THREATS TO EXTERNAL VALIDITY

The concept of **external validity,** which some argue was originally confounded with other types of validity, is currently defined as the "validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables" (Shadish et al., 2002, p. 38). Nonetheless, external validity is often used by researchers as a synonym for generalizability or representativeness. Lynch (1982) identified three issues that are frequently conflated in this broader use of the term. First, there is *statistical generalizability,* which refers to the representativeness of the results to a wider population of interest. Second is *conceptual replicability or robustness,* which Lynch believed to be closest to Campbell and Stanley's (1966) conception of external validity. Third is *realism,* which is also similar to what Aronson and Carlsmith (1968) referred to as **mundane realism,** or the extent to which an experimental treatment is apt to be encountered in an analogous form in a natural setting. Incidentally, Aronson and Carlsmith made a further distinction between mundane and **experimental realism,** the latter referring to the psychological impact of the experimental manipulation on the participants. As if external validity were not already elusive enough, Lynch (1982) also argued that it "cannot be evaluated either a priori or a posteriori (e.g., on the basis of sampling practices or realism) in the absence of a fairly deep understanding of the structural determinants of the behavior under study" (p. 239). Lynch's position is that we must have an implicit or explicit model of the behavior we are investigating, or else we leave judgments of the external validity of experiments to experts in the substantive area who have a sense of the behavior under investigation (i.e., as opposed to mere specialists in methodology).

If external validity seems a mercurial concept and not easy to pin down, another issue is that a number of psychological experimenters have questioned the importance of external validity as a criterion of a sound experimental design. Douglas Mook (1983) contended that the insistence on external validity is often misguided. The point of many laboratory simulations, he argued, is not to generalize to the real world, but instead to try to make predictions about the real world from the laboratory. Mook noted the monkey love studies of Harlow (discussed in chapter 1) as an example of experiments that are lacking in external validity (because using baby monkeys, and wire mesh or cloth-covered mother surrogates, to study human babies falls far short of the ideal) but that nevertheless tell us something theoretically valuable about personality development. Mook cautioned that, before condemning any experiment as lacking in external validity, it would be far more instructive to ask: (a) Is the investigator

really trying to estimate from sample characteristics the characteristics of a population, or is the purpose of the study instead to draw conclusions about a theory that predicts what *these* subjects will do? And (b) Is the purpose of the study to predict what would happen in a real-life situation, or is its purpose to test under very controlled conditions a predicted causal relationship that is purported to be a universal principle of behavior (and that should therefore operate in the laboratory as well as in real life)? More recently, Anderson, Lindsay, and Bushman (1999) reexamined the issue addressed by Mook, this time by asking whether there is actually a correspondence between lab and field experimental findings that have addressed similar questions. They inspected the effect sizes in a range of studies (including studies of aggression, helping, leadership style, social loafing, self-efficacy, depression, and memory) and concluded that there was considerable similarity in the pooled effect sizes of laboratory and field studies using conceptually similar independent and dependent variables.

Some years before the terms *external validity* and *internal validity* were coined by Campbell and Stanley, another noted experimentalist, Egon Brunswik (1947), had addressed the issue of representativeness in a way that now seems conceptually related to both external and internal validity. If we want to generalize the results of a psychological experiment to a population of subjects and a population of stimuli, then we must sample from both populations, Brunswik argued (see also discussion by Maher, 1978). Brunswik used the expression **representative research design** to describe an idealized experimental model in which both the subjects and the experimental stimuli are representative of specified populations. Experiments that satisfy this criterion were called **ecologically valid.** Suppose we wanted to test the hypothesis that male and female patients respond quite differently to a certain psychotherapeutic treatment when the clinician is male or female. A convenient experimental design would consist of randomly assigning patients of both genders to either a male or a female clinician. Though it might be claimed that the design is representative in terms of the selection of patients (assuming they were properly sampled), it could *not* be claimed that the design is representative as regards the stimulus (i.e., the clinicians presenting the treatments). Because the experimenter did not sample from among populations of male and female clinicians, we would be hard-pressed to conclude that there is a generalizable relationship of the type hypothesized. Thus, Brunswik might say that, inasmuch as the use of other male or female clinicians might produce quite different results, the design of the study is deficient in ecological validity (Hammond, 1954)—and by implication, we might add, deficient in external validity as well.

What does this also have to do with internal validity? The argument is that our use of only one clinician of each sex as a stimulus does not preclude the possibility that some other characteristics of this person may have stimulus values that are unknown and uncontrolled for. In other words, there are two major limitations in this so-called "single stimulus design" (Maher, 1978). First, it is possible that differences among the patients who are exposed to the male clinician and those exposed to the female clinician may be due to the effects of uncontrolled stimulus variables. On the basis only of the information available from our data, we cannot conclude whether the obtained differences are due to the validity of the tested hypothesis or to the effects of another uncontrolled variable (i.e., clearly a threat to internal validity). Second, the failure to find differences between those subjects exposed to the male clinician and those exposed to the female

clinician might also be due to the presence of an uncontrolled stimulus variable operating either (a) to counteract the effect of the intended independent variable or (b) to increase that effect artificially to a ceiling value (i.e., a top limit) in the different groups. We have no way of distinguishing between this explanation and the possibility that the lack of difference is due to the invalidity of the tested hypothesis.

Earlier we mentioned the idea of moderator variables that affect relationships between independent and dependent variables. Presumably, given an adequate theory, we can formulate a model on which to predicate the carving out of moderator variables. For example, Alice H. Eagly (1978) was intrigued by the claim that, generally speaking, women are more conforming and more easily influenced than men. The explanation proposed for this idea was that socialization processes taught men to be independent thinkers, a cultural value not as frequently thought to be suitable for women. The empirical findings, however, were inconsistent. Some failed to find gender differences in influenceability. Using a historical model in which she reasoned that the era in which the empirical data were collected was a plausible moderator of the association between gender and influenceability, Eagly meta-analyzed all the relevant studies she could find. Just as her model predicted, there was a pronounced difference in the correlation between gender and influenceability in studies published before 1970 and those published during the era of the women's movement in the 1970s. In contrast to the older research studies, which had found greater influenceability among females than among males, the later studies identified few gender differences in influenceability.

Brinberg et al. (1992) cautioned that when researchers know little about the moderator variables lurking beneath the surface of their aggregated variables, they may unwittingly misrepresent the external validity of their causal inferences and recommendations based on those inferences. It is quite possible, for example, that critical patterns that hold true in the aggregate may not hold for only a small number of individuals (Hutchinson, Kamakura, & Lynch, 2000; Yule, 1903), and thus, it is always prudent to explore the individual data. In biomedical research, standard moderator variables include demographic descriptors like age, sex, ethnic group, and prior pathology. Suppose that research shows a particular treatment of flu is effective. Still, we want to break down the aggregate scores so that we can state with more precision when the treatment can be expected to be *most* (and *least*) effective. We might find that Caucasian men do better on certain dosages of the treatment than non-Caucasian men, or that both men and women with prior pathology do the poorest, or that younger people do better than older people in some ethnic groups. In the field of experimental psychology it is quite common for researchers to rely on what are called *convenience samples,* which simply means samples made up of people who are readily accessible, usually sophomores in introductory psychology courses. As a way of exploring for possible moderator variables, it is standard practice in many psychology departments to request that students in introductory psychology classes complete a battery of psychological instruments (typically including some measure of major factors of individual personality; cf. Goldberg, 1993; McCrae & Costa, 1997; Wiggins, 1996), as well as provide demographic information that can then be correlated with the students' total scores or with residuals about the mean in the research in which they participate.

Frequently, the problem with convenience samples is that researchers seem oblivious even to the possibility that their subject samples may not be representative

of the population they are presumed to be theorizing about, as if all humans were the same, or all rats were the same. We return to this issue in chapter 9, but the latter problem was illustrated some years ago by Marshall B. Jones and Robert S. Fennell, III (1965). In the 1940s, there was a controversy between two leading psychological researchers, Clark L. Hull and Edward C. Tolman, over the nature of learning. Hull, inspired by Pavlov's research on conditioned reflexes, developed a systematic behavior theory that asserted that animals learned stimulus-response connections and that the strength of these connections accumulated in small increments from trial to trial. In contrast, Tolman's model (known as *purposive behaviorism, sign gestalt theory,* or *expectancy theory*) emphasized the cognitive nature of learning, arguing that animals learned "what leads to what" by acquiring expectations and forming "cognitive maps." Learning is not an automatic, mechanical process, but a discontinuous process that largely depends on exploratory behaviors, the Tolmanians argued.

Not only were there distinct theoretical and methodological differences between those two camps, but they also used different strains of rats. The Tolmanians, centered at the University of California, used rats that had been selectively descended from crossed matings of wild males and laboratory albino females. The Hullians, at Yale University under Hull's direction and a second camp at Iowa University under Kenneth W. Spence, used descendents of a "nonemotional" strain of rats that had descended from very different crossed matings. That the two strains of rats had been separated for over thirty years, during which time they had been differently and selectively bred, raised the question of whether genetic differences were involved in the clearly different results of the Hullians and the Tolmanians. So Jones and Fennell (1965) obtained a sample of rats from each strain, placed them on a 23-hour food or water deprivation schedule and, beginning on the fourth day, subjected them to three learning trials daily in a U-maze for ten consecutive days. There were noticeable differences in the performance of the two strains, differences that were also entirely consistent with the nature of the theoretical differences that separated the two schools of learning. The Tolman rats "spent long periods of time in exploratory behaviors, sniffing along the walls, in the air, along the runway" (p. 294). In contrast, the Hull-Spence rats "popped out of the start box, ambled down the runway, around the turn, and into the goal box" (p. 294). Findings like these would not necessarily lead us to question either the logic or the internal consistency of Hull's or Tolman's theory of learning, but they do raise a serious question about the external validity of the empirically based causal generalizations that "were involved in the great debate over the nature of learning" (p. 295).

STATISTICAL CONCLUSION AND CONSTRUCT VALIDITY

Besides internal and external validity, there is statistical conclusion validity and construct validity. As defined by Shadish et al. (2002), **statistical conclusion validity** is concerned with "inferences about the correlation (covariation) between treatment and outcome" (p. 38), in other words, Hume's "contiguity of events." If we are interested in effect sizes, for example, the question of interest is whether a statement about the association between membership in the treatment or control group and the dependent variable can be made

with a reasonable degree of confidence. If we are using a significance test, was there enough statistical power to detect a likely relation between the treatment and outcome and to rule out the possibility that the observed association was due to chance? In the second half of this book we have more to say about statistical power, assumptions of particular tests of statistical significance, and related issues. Among the threats to statistical conclusion validity discussed by Shadish et al. are low statistical power (in which case we are apt to make Type II errors), violations of assumptions of statistical tests (which lead to spurious estimates of p values), "fishing" for statistically significant effects without making proper adjustments of p values, unreliable tests and measurements, and spurious or uninterpretable or ambiguous estimates of effect sizes.

Turning finally to construct validity, recall our discussion in chapter 4, where we defined *construct validity* as referring to the degree to which a test or questionnaire measures the characteristic that it is presumed to measure. We also noted that, as Popper's falsificationist view implies, constructs can never be completely verified or proved, because it is impossible to complete every conceivable check on the construct (Cronbach & Quirk, 1971). Shadish et al. (2002) define **construct validity** as referring to "higher order constructs that represent sampling particulars" (p. 38). In research in which causal generalizations are the prime objective, *construct validity* is the soundness or logical tenability of the hypothetical idea linking the independent (X) and dependent (Y) variables, but it also refers to the conceptualization of X and Y. One way to distinguish between construct validity and internal validity is to recall that internal validity is the ability to logically rule out competing explanations for the observed covariation between the presumed independent variable (X) and its effect on the dependent variable (Y). Construct validity, on the other hand, is the validity of the theoretical concepts we use in our measurements and causal explanations. Thus, whenever we ask what is *really* being measured (e.g., "What does this test really measure?"), we are asking about construct validity rather than internal validity.

Put another way, construct validity is based on the proper identification of the concepts being measured or manipulated (i.e., "Do we have a clear conception of what we are measuring or manipulating?"), and internal validity is based on whether a variable other than X (the causal variable we *think* we are studying) may have caused Y to occur. Hall (1984a) proposed a further distinction among the four kinds of validity in experimental research. Poor construct or internal validity has the potential to actively mislead researchers because they are apt to make causal inferences that are plain "wrong." Poor statistical-conclusion or external validity puts the researchers in a "weak position" to make *any* causal inferences or broad generalizations, because it limits what can be learned or what can be generalized to other situations. Thus, according to Hall's argument, the distinction comes down to being wide of the mark (i.e., poor construct or internal invalidity) or being in a vulnerable position on statistical or sampling grounds (statistical-conclusion and external validity).

SUBJECT AND EXPERIMENTER ARTIFACTS

We turn now to a class of threats to the construct, internal, and external validity of experiments (as well as threats to the valid interpretation and generalization of nonexperimental results) that we have studied and written about for many years

(e.g., Rosenthal, 1966, 1976; Rosenthal & Rosnow, 1969a; Rosnow & Rosenthal, 1997). The term **artifacts** is used, generally, to refer to research findings that result from factors other than the ones intended by the researchers, usually factors that are quite extraneous to the purpose of their investigations (e.g., Orne, 1959; Rosenthal & Rosnow, 1969a; Rosnow, Strohmetz, & Aditya, 2002). By **subject and experimenter artifacts,** we mean that systematic errors are attributed to uncontrolled subject- or experimenter-related variables (Rosnow, 2002). The term *experimenter* is understood to embrace not only researchers who perform laboratory or field experiments, but those working in any area of empirical research, including human and animal experimental and observational studies. The sociologist Herbert H. Hyman and his colleagues (1954) wisely cautioned researchers not to equate ignorance of error with lack of error, because all scientific investigation is subject to both random and systematic error. It is particularly important, they advised, not only to expose the sources of systematic error in order to control for them, but also to estimate the direction (and, if possible, the magnitude) of this error when it occurs. The more researchers know about the nature of subject and experimenter artifacts, the better able they should be to isolate and quantify these errors, take them into account when interpreting their results, and eliminate them when possible.

Though the term *artifact* (used in this way) is of modern vintage, the suspicion that uncontrolled sources of subject and experimenter artifacts might be lurking in investigative procedures goes back almost to the very beginning of modern psychology (Suls & Rosnow, 1988). A famous case around the turn of the twentieth century involved not human subjects, but a horse called Clever Hans, which was reputed to perform remarkable "intellectual" feats. There were earlier reports of learned animals, going all the way back to the Byzantine Empire when it was ruled by Justinian (A.D. 483–565), but no animal intelligence captured the imagination of the European public and scholars alike as that attributed to Hans (Rosenthal, in Pfungst, 1965). Hans gave every evidence that he could tap out the answers to mathematical problems or the date of any day mentioned, aided ostensibly by a code table in front of him based on a code taught to him by his owner. It seemed unlikely that his owner had any fraudulent intent because he allowed visitors (even in his absence) to question Hans, and he did not profit financially from the horse's talents. Thus, it was possible to rule out intentional cues as the reason for the horse's cleverness. One visitor, the German psychologist Oskar Pfungst, discovered in an elegant series of experiments that Hans's accuracy diminished when he was fitted with blinders so he could not see his questioners, when the distance between Hans and his questioners was increased, or when the questioner did not know the answer. These results implied that the horse's apparent talents were due to something other than his capacity to reason. Pfungst found that Hans was responding to subtle cues given by his questioners, not just intentional cues, but unwitting movements and mannerisms (Pfungst, 1911). For instance, someone who asked Hans a question that required a long tapping response would lean forward as if settling in for a long wait. The horse responded to the questioner's forward movement, not to the actual question, and kept tapping away until the questioner unconsciously communicated the expectancy that Hans would stop tapping. This the questioner might do by beginning to straighten up in anticipation that Hans was about to reach the correct number of taps.

Pfungst's unraveling of the mystery of Clever Hans provided an object lesson in the susceptibility of behavior (even animal behavior) to unconscious suggestion. Given the influence on animal subjects, might not the same phenomenon hold for human subjects who are interacting with researchers oriented by their own hypotheses, theories, hunches, and expectations? Although Pfungst's discovery was duly cited and circulated, its wider methodological implications did not strike a resonant chord in behavioral science during that period. To be sure, a number of leading experimenters, including Hermann Ebbinghaus (1885, 1913), voiced their suspicions that researchers might unwittingly influence their subjects. However, their concerns, along with the wider methodological implications of Pfungst's discovery, went largely unheeded until, several decades later, another influential development fostered the idea that human subjects behave in special ways *because* they know they are "subjects" of an investigation. This principle, which came to be known as the **Hawthorne effect,** grew out of a series of human factors experiments between 1924 and 1932 by a group of industrial researchers at the Hawthorne works of the Western Electric Company in Cicero, Illinois (Roethlisberger & Dickson, 1939). One set of studies examined the impact of higher levels of electric lighting, increased rest periods, and other conditions on the work productivity of young women who inspected parts, assembled relays, or wound coils (Gillespie, 1988). According to news reports and a Western Electric memorandum, one study revealed that any improvement in working conditions resulted in greater worker satisfaction and increased productivity. When the improvements were removed, however, the productivity did not decline; the efficiency actually continued to increase, according to the reports. On interviewing the team of six workers who had participated in that study, the researchers concluded that the workers' productivity increases had derived from their feeling flattered by being subjects of investigation. That is, they had been motivated to increase their output because of their special status as research participants. Not only had their opinions been solicited by management, but they had been singled out for free morning tea, rest periods, and shorter hours of work.

The term *Hawthorne effect* was coined by the contributor of a chapter to a popular textbook in the 1950s (French, 1953). Subsequently, however, the original reports and secondary accounts of this study were subjected to critical analysis by other investigators (cf. Adair, 1984; Bramel & Friend, 1981; Franke & Kaul, 1978; Gillespie, 1988; Schlaifer, 1980), who argued, among other things, that the historical record was tainted by sweeping generalizations embroidered by overzealous, and possibly biased, authors. In another fascinating piece of detective work, H. McIlvaine Parsons, a specialist in human factors research, described his discovery of a long-ignored confounding variable that also explained the Hawthorne effect (Parsons, 1978, 1992). The assembly-line workers in the Hawthorne studies had been told their output rates, and the higher the rates, the more they were paid, Parsons learned. Putting those facts together, he theorized that the increased productivity had been reinforced by the feedback the workers had received about their output rates. Like some projective test into which people read their own meanings, the Hawthorne effect was a mixture of fantasy and reality into which textbook authors had read their own meaning, Parsons argued. Nevertheless, the principle of the Hawthorne effect entered into the vocabulary of behavioral research as implying a kind of "placebo effect" in psychological research

with human subjects (Sommer, 1968). That is, it implies that subjects responded not just to the experimental treatment, but also to uncontrolled factors, including the belief that they were being administered a treatment designed to have a particular effect. A generation of researchers was warned to be wary of unleashing a Hawthorne effect by their manipulations, observations, or measurements.

In 1933, another important development (which went largely unnoticed for many years) involved a conceptual advance. Saul Rosenzweig, a clinical psychologist fresh out of graduate school, published an insightful critique in a leading psychology journal, in which he examined various aspects of the psychology experiment and identified three distinct sources of artifacts. For example, he described how artifacts might result from, first, the "observational attitude" of the experimenter. Using chemistry as his prototype of scientific experimentation, he noted how chemists take into account the ambient temperature, possibly even their own body heat, when running certain lab experiments. Experimenting psychologists, Rosenzweig said, needed to consider their own attitudes toward their research subjects and the subjects' beliefs about and attitudes toward the experiment. His second point was that, of course, "chemicals have no power of self-determination" (p. 338), whereas experimenting psychologists usually work with people who may try to outguess the experimenter and to figure out how their behavior will be evaluated. Rosenzweig called this the "motivational attitude" problem, and he claimed that it could creep into any experiment and bias the results. Third were what he called "errors of personality influence," for example, the warmth or coolness of the experimenter, his or her unguarded gestures or words, and the experimenter's sex and race—all possible confounding factors that might affect the attitudes and reactions of the research subjects, quite apart from the experimental treatment. Rosenzweig sketched some procedures that he thought would prevent some of these problems, including the use of deceptions to prevent errors of motivational attitude. Nonetheless, he also cautioned that it was frequently unclear whether the experimenter or the subject was the "true deceiver"—a concern voiced again by other writers in the 1960s (e.g., Stricker, 1967).

Beginning in the 1960s and throughout the 1970s, increased attention was paid to concerns about subject and experimenter artifacts. There are several possible reasons for that development, one of which has to do with the rise of cognitive psychology. First, many earlier behavioral psychologists had been fixed on a dustbowl-empiricist view that emphasized only observable responses as acceptable data in science, but renewed interest in the cognitive dimension and the neobehaviorist reshaping of the empiricist view made it respectable to talk about cognition as a variable of scientific relevance (Toulmin & Leary, 1985). A second reason was that scientific psychology was coming into its own; its identity crisis seemed virtually over (Silverman, 1977, pp. 18–19). Following World War II, there had been a tremendous growth in psychology departments and an increased role for research psychologists in the government, the military, and industry as a result of optimism about the likely benefits of psychological science. Those who voiced concern over artifacts were seen as undermining the empirical foundations of the scientific facts and theories that were proliferating in psychological science. By the 1960s and 1970s, increasing numbers of researchers felt secure enough to, as Hyman (1954, quoted earlier) put it, accept

that ignorance of error was not synonymous with lack of error, but that it merely signaled a primitive understanding and thus a less advanced stage of scientific development. For these, and probably other good reasons, the stage was set for programmatic investigations of subject and experimenter artifacts by researchers working independently in different institutions.

DEMAND CHARACTERISTICS AND THEIR CONTROL

Among those in the first wave of contemporary artifact researchers was Martin T. Orne, an eminent psychiatrist, social psychologist, and clinical psychologist at the University of Pennsylvania. Starting in the late 1950s, Orne had begun to explore the role of uncontrolled task-orienting cues in experimental research. He was primarily interested in the complex nature of hypnosis when he began this program of investigation and had observed that, at the conclusion of many of his hypnosis experiments, the subjects asked questions such as "Did I ruin the study?" By the use of sensitive postexperimental interviewing, he learned that what the subjects were asking was "Did I perform well in my role as experimental subject?" or "Did my behavior demonstrate what the study was designed to show?" That is, it appeared that the subjects were responding, at least in part, to what they interpreted as cues about what the experiment was "really" about and what the experimenter "wanted" to find out. Borrowing a concept from the theoretical work of Kurt Lewin (1935)—*Aufforderungscharakter* (or "demand value")—Orne (1959) coined the term **demand characteristics** to denote the subtle, uncontrolled task-orienting cues in an experimental situation. In an earlier paper, Sarbin (1944) had drawn an analogy with the Heisenberg effect in atomic physics to argue that the observation or measurement of behavior could alter the behavior observed. In fact, a similar idea had been anticipated by Rosenzweig (1933), whose work we discussed above, particularly his "motivational attitude" idea. Orne and his associates advanced this idea a giant step by demonstrating, in a series of ingenious studies, how demand characteristics could produce artifacts in the research.

In one early study, using college students in an introductory psychology course as the participants, Orne (1959) conducted a demonstration of hypnosis on several subjects. The demonstration subjects in one section of students were given the suggestion that upon entering a hypnotic trance, they would manifest "catalepsy of the dominant hand." All the students in this section were told that catalepsy of the dominant hand was a standard reaction of the hypnotized person, and the group's attention was called to the fact that the right-handed subject had catalepsy of the right hand and the left-handed subject had catalepsy of the left hand. In another section (the control group), the demonstration of hypnosis was carried out, but without a display of Orne's concocted "catalepsy" reaction. In the next phase of the study, Orne asked for volunteers for hypnosis from each section and, after they had been hypnotized, had them tested in such a way that the experimenter could not tell which lecture they had attended until after the completion of the experiment. Of the nine volunteers from the first section (the one in which catalepsy of the dominant hand had been demonstrated), five of them showed catalepsy of the dominant hand, two showed catalepsy of both hands, and two showed no catalepsy. None of the nine control

subjects showed catalepsy of the dominant hand, but three of them showed catalepsy of both hands. Because catalepsy of the dominant hand (the reaction that Orne had invented) was known not to occur spontaneously, its occurrence in the first group but not in the second was interpreted by Orne as support for his demand characteristics theory. That three of the nine subjects in the control group spontaneously displayed catalepsy of both hands was explained by him in terms of the experimenters' repeated testing for this reaction, which Orne thought may have introduced its own set of implicit demand cues.

Orne referred to this cooperative behavior as the **good subject effect,** and he argued that subjects would often go to remarkable lengths to comply with demand characteristics. For example, at one point in his research on hypnosis he tried to devise a set of dull, meaningless tasks that participants who were not hypnotized would refuse to do or would try for only a short time and then abandon. One task consisted of adding hundreds of thousands of two-digit numbers. Five and a half hours after the subjects began, Orne gave up! Even when the subjects were told to tear each worksheet into a minimum of 32 pieces before going on to the next, they persisted in adding up the digits. Orne explained this behavior as the role enactment of volunteer subjects who reason that, no matter how trivial and inane the experimental task seems to them, it must surely have some important scientific purpose or they would not have been asked to participate in the first place. Thus, he theorized, they complied with the demand characteristics of the experiment in order to "further the cause of science" (Orne, 1962).

Orne gained another insight into the good subject effect when he asked a number of casual acquaintances to do an experimenter a favor and, on their acquiescence, asked them to do five push-ups. They seemed amazed and incredulous, and all responded "Why?" When he asked a similar group of individuals whether they would take part in an experiment and, on their acquiescence, asked them to do five push-ups, their typical response was "Where?" (Orne, 1962). What could account for the dramatic difference in responses? Orne theorized that people who agree to participate in an experiment implicitly agree to comply with whatever demand cues seem implicit in the experimental situation. Subjects are concerned about the outcome of the experiment in which they have agreed to participate. Consequently, they are motivated to play the role of the good subject who responds to overt and implicit cues in ways designed to validate the experimenter's hypothesis. Other researchers obtained similar kinds of effects, all suggesting compliance with demand characteristics. The phenomenon also seemed wide-ranging, as it was demonstrated in attitude change research, prisoners' dilemma games, verbal operant conditioning, testing, and on and on (for further discussion and citations, see Rosnow & Rosenthal, 1997, p. 68). Furthermore, Orne surmised, it is not possible to control for the good subject effect in the classic sense; what is needed is a means of ferreting out the demand characteristics in each experimental situation. In theory, he thought, having this information should allow researchers to interpret their data more accurately and, sometimes, even to circumvent the demand artifact in question.

Guided by this vision, Orne proposed that researchers use the subjects themselves to assist in the detection and interpretation of demand characteristics. It is important, he argued, not to attribute to demand characteristics even more potency

than they possess, for that will surely lead to "a nihilistic view at least as naïve as that which denies the potential importance of these factors" (Orne, 1970, p. 260). In science, the proof of the pudding is in confronting a problem empirically, and thus, Orne showed that it was no longer necessary merely to speculate on the role of "errors of motivational attitude." In what he called a **quasi-control strategy,** his idea was to have some of the research subjects step out of the good subject role and act as "coinvestigators" in the search for truth. Orne proposed several techniques for having quasi-control subjects reflect on the experiment and tell how their behavior might be compromised or influenced by uncontrolled factors rather than by the controlled independent variable. One technique was to have subjects serve as their own quasi controls in postexperimental interviews. In these interviews, they were asked to disclose the factors that were important in determining their reactions in the experiment and to reveal their beliefs about, and perceptions of, the experiment and the experimenter. These subjects must be convinced that the study is over and that they are now playing the role of coinvestigators (or aides), or the data they provide should also be suspect as biased by demand characteristics.

In another use of quasi controls, called *preinquiry* by Orne, some of the prospective subjects are sampled and afterward are separated from the subject pool. The experimental procedures are then carefully described to these quasi controls, and they are asked to speculate on how they would be likely to behave in the experiment. Comparisons are later made between their projected role responses and the actual responses of the participating subjects. In this way, Orne theorized, it should be possible to get an insight into how the experimental outcome might be affected by the real subjects' guesses and role responses to how they *should* behave. Still another alternative used what Orne called a "sacrifice group" of quasi controls. These are people who are pulled out of the experiment at different points and questioned about their perceptions of the experiment up to that point. Another option discussed by others is to have the preinquiry individuals tell how they think they would react to different deception treatments. The idea here is that, if no differences are apparent between different intensities of deception, the least intense deception should be as effective as the most intense deception (Fisher & Fyrberg, 1994; Suls & Rosnow, 1981).

Orne noted the volunteer status of his research subjects, as well as the fact that they seemed to be remarkably cooperative. Insights like these inspired other researchers to compare volunteer subjects and nonvolunteer subjects (e.g., coerced participants or captive participants) on a range of tasks (Rosenthal & Rosnow, 1975), using volunteer status as a proxy for the "good subject." Horowitz (1969) observed that volunteers responded differently from nonvolunteers to fear-arousing communications in an attitude change experiment. In our earlier discussion of the Solomon design, we mentioned the finding that the volunteer status of subjects was also associated with reversals in pretest-treatment interactions in an attitude change experiment, the volunteers again being the more compliant participants (Rosnow & Suls, 1970). Kotses, Glaus, and Fisher (1974) reported volunteer biases in a study of physiological responses to random bursts of white noise, and Black, Schumpert, and Welch (1972) observed that perceptual-motor responses were also associated with subjects' volunteer status. In another study, the volunteer status of the participants in a verbal operant-conditioning study was associated with a greater degree of compliance with demand cues (Goldstein,

Rosnow, Goodstadt, & Suls, 1972). Volunteer bias has also been found in clinical and counseling studies (King & King, 1991; Strohmetz, Alterman, & Walter, 1990). We will have more to say about volunteer subjects and strategies for predicting the direction of the response bias in chapter 9, but it appears that volunteers for research participation tend to be more sensitive and accommodating to demand cues than are coerced subjects or captive nonvolunteers.

Interestingly, not all artifact researchers agreed with Orne's conception of the good subject effect. For example, Milton J. Rosenberg got into a spat with some leading cognitive dissonance theorists when he argued that an experimental design used in an important cognitive dissonance study had produced spurious effects resulting from the participants' anxieties about how they would be evaluated—which Rosenberg (1965) called **evaluation apprehension.** Based on a series of other experiments of his own, he found that when subjects worry that the experimenter plans to evaluate an aspect of their performance, they behave in ways designed to win the experimenter's approval or to avoid disapproval. Experiments in which some evaluation apprehension appeared likely were those containing an element of surprise or having an aura of mystery to them. The more explicit the cues, the more control the experimenter has in granting positive evaluation, and the less effortful the subjects' responses, the greater may be the resulting response bias due to the subjects' feelings of evaluation apprehension. One solution to this problem may be to ensure the confidentiality of the subjects' responses, on the assumption that individual subjects will then be less apprehensive and more forthcoming in their responses (e.g., Esposito, Agard, & Rosnow, 1984). However, in some research—for example, research on sensitive topics (such as sexual behavior and AIDS)—it may be exceedingly difficult to control for evaluation apprehension and related problems (e.g., Catania, Gibson, Chitwood, & Coates, 1990). It is also conceivable that in some (probably rare) experimental situations some subjects may feel a conflict between evaluation apprehension and the good subject effect (e.g., Rosnow, Goodstadt, Suls, & Gitter, 1973; Sigall, Aronson, & Van Hoose, 1970), in which case the evidence suggests that "looking good" may emerge as the predominant motivation of many subjects, as opposed to helping the cause of science (i.e., "doing good").

INTERACTIONAL EXPERIMENTER EFFECTS

In chapter 5 we spoke of noninteractional artifacts, that is, artifacts that are not directly associated with the interaction between the experimenter and the research subjects. Two general classes discussed in that chapter were interpreter and observer biases. The other side of this coin comprises five general classes of artifacts called **interactional experimenter effects** (Rosenthal, 1966, 1976). These artifacts are recognized by being attributable to some aspect of the interaction between experimenters and their subjects. We first briefly describe all five of these classes (i.e., biosocial attributes, psychosocial attributes, situational factors, modeling effects, and expectancy effects) and then discuss the fifth type and its control in greater detail. Researchers interested in learning more about the nature and control of subject and experimenter artifacts will find a fully detailed discussion of experimenter effects in Rosenthal's (1966, 1976) *Experimenter Effects in Behavioral Research* and a more

recent theoretical and ethical overview in our book entitled *People Studying People* (Rosnow & Rosenthal, 1997).

First, *biosocial attributes* include the biological and social characteristics of experimenters, such as gender, age, and race. For example, a good deal of research has been reported showing that male and female experimenters sometimes obtain significantly different data from their subjects. It is not always possible to predict for any given type of experiment just how subjects' responses will be affected by the experimenter's gender, if indeed there is any effect at all. However, when such effects have occurred, it seems that the male and female experimenters behaved differently toward their subjects, thereby eliciting different responses because the experimenters had altered the experimental situation for the subjects (e.g., Barnes & Rosenthal, 1985). In one study, the male experimenters were found to be friendlier than the female experimenters. It was also found that 12% of the experimenters, overall, smiled at their male subjects, whereas 70% smiled at their female subjects (Rosenthal, 1967, 1976). A further finding was that smiling by the experimenters predicted the results. The lesson is that before we claim a gender difference in the results of behavioral research, we must make sure that male and female subjects were treated identically. If they were not, then gender differences in the results might be due not to constitutional or socialization variables, but to the fact that male and female subjects did not participate in the "same" experiment (i.e., they were treated differently).

Whereas biosocial attributes are usually readily accessible by inspection, the second class, termed *psychosocial attributes,* are readily accessible but not simply by inspection. These attributes include factors such as personality and temperament, which are often assessed more indirectly, frequently by the use of standard psychological tests or trained observers' judgments. For example, experimenters who differ in anxiety, approval need, hostility, authoritarianism, status, or warmth also tend to obtain different responses from their subjects. Experimenters higher in status generally have a tendency to elicit more conforming but less pleasant responses from their subjects, and experimenters who are warmer in their interactions with the subjects often obtain more competent and more pleasant responses. Examiners who act more warmly to people being administered a test of intelligence are apt to elicit better intellectual performance than are cooler examiners or examiners who are perceived as threatening. In simple tasks with ostensibly little meaning, the subjects' expectations may assume increasingly greater importance. The subjects who view experimenters more favorably may view the tasks more favorably, thus transforming a compellingly inane procedure into one that simply "must" have more value. An experimenter perceived as threatening might arouse feelings of evaluation apprehension, leading to a more defensive posture or simply distracting the subjects from the task and thus eliciting less-than-ideal performance.

Third are *situational effects*. More than experimenters' scores on a psychological test of anxiety or approval need, their status and warmth are defined and determined in part by the nature of the experimental situation and the particular subject being contacted. Experimenters who are acquainted with their subjects may behave differently toward them than toward unfamiliar subjects. Experimenters who are more experienced in conducting a given experiment often obtain different responses from subjects than

do less experienced experimenters. Things that happen to experimenters during the course of their experiments, including the responses they obtain from their first few subjects, may also influence the experimenters' behavior, and in turn, those changes may lead to further changes in subjects' responses. When the first few subjects respond as they are expected to respond, the behavior of the experimenter may change in such a way as to influence the subsequent subjects to respond too often in the direction of the experimenter's hypothesis (Rosenthal, 1976). Thus, when subjects are run one at a time, we may want to block on (subdivide by) time periods, to see whether the results are similar at the beginning, middle, and end of the experimental trials.

A fourth type of interactional experimenter artifact is a **modeling effect.** It sometimes happens that before the experimenters conduct their studies, they try out the tasks that they will later have their subjects engage in. Although the evidence on this point is not all that clear, it would seem that, at least sometimes, the investigator's own performance becomes a factor in the subjects' performance. When the experimental stimuli are ambiguous, subjects' interpretations of their meaning may too often agree with the investigator's interpretations of the stimuli. The problem is that the experimenter's behavior, rather than the hypothesized psychological processes, may have produced the results (Rosenthal, 1976). In survey research, there is evidence that the interviewer's own opinion, attitude, or ideology may affect the responses obtained from the respondents. If a modeling effect occurs, it is most likely to be patterned on the interviewer's opinion or attitude, but in a minority of cases the subjects may respond in a direction opposite to that favored by the interviewer (Rosenthal, 1976). In laboratory studies, it appears there is a tendency for happier, affable, less tense experimenters to model their subjects negatively, and for less pleasant, more tense experimenters to model their subjects positively. Why this should be so is unclear, but one methodological implication may be to use more naturally "neutral" experimenters in order to reduce the possibility of modeling effects.

Generally speaking, the most critical control for all four classes of interactional artifacts above is woven into the fabric of science by the tradition of replication. This is also true of a fifth type of artifact, experimenter expectancy, but there are other ways of addressing this particular problem (which we discuss in the next section). The term *experimenter expectancy* takes its name from the idea that some expectation of how the research will turn out is virtually a constant in science. In the same way that the questioners of Clever Hans unintentionally altered their own behavior and that in turn affected the horse's responses, so can hypotheses, theories, or expectations that are held by experimenters lead them unintentionally to alter their behavior toward their subjects. We are speaking, then, of the investigator's hypothesis or expectancy as a **self-fulfilling prophecy,** but not exactly in the way this term was conceived of by its originator, Robert Merton (1948), who defined it as a "*false* definition of the situation evoking a new behavior which makes the originally false conception come *true*" (p. 195). By **experimenter expectancy effect,** we mean that the experimenter's expectation (true *or* false) may come to serve as a self-fulfilling prophecy, which can be conceived of as a type of interpersonal expectancy effect. That is, someone acting in accordance with a personal set of expectations treats another individual in such a manner as to increase the likelihood of eliciting behavior that conforms to the first person's expectations (e.g., Blanck, 1993). An example

would be a teacher who believes certain pupils are especially bright and then acts toward these pupils more warmly, teaches them more material, and spends more time with them, behavior that, over time, results in greater gains in achievement for these students than would have occurred in the absence of the interpersonal expectation (Rosenthal & Jacobson, 1968).

EXPERIMENTER EXPECTANCY EFFECTS AND THEIR CONTROL

In one early study designed to demonstrate the effects of experimenters' expectancies on the results of their research, the experimenters were given rats that were to be taught to run a maze with the aid of visual cues (Rosenthal & Fode, 1963). Half the experimenters were told their rats had been specifically bred for maze brightness, and the remaining experimenters were told their rats had been bred for maze dullness. Actually, there were no differences between the rats assigned at random to each of the two groups. At the end of the experiment the results were clear. The rats run by the experimenters expecting brighter behavior showed learning significantly superior to that of the rats run by the experimenters expecting dull behavior. The study was later repeated, this time using a series of learning trials, each conducted in Skinner boxes (Rosenthal & Lawson, 1964). Half the experimenters were led to believe their rats were "Skinner box bright"; the other experimenters were led to believe their animals were "Skinner box dull." Once again, there were not really any differences in the two groups of rats, at least not until the results were analyzed at the end of the study. Then, the allegedly brighter rats really did perform better, and the alleged dullards really did perform more poorly. Neither of the animal studies showed any evidence that the student experimenters might have been falsifying their results. Thus, it could be concluded that the experimenters' expectations had acted not on the experimenters' evaluation of the animals' performance, but on the actual performance of the rats.

In the period since those two studies were conducted, literally hundreds of additional studies have examined the possible occurrence of expectancy effects both inside and outside the experimental lab (e.g., Harris & Rosenthal, 1985; Rosenthal & Rubin, 1978). By the beginning of the 1990s, there were over 450 studies. In a meta-analysis of 345 studies in the 1970s (Rosenthal & Rubin, 1978), the probability of no relation between experimenters' expectations and their subjects' subsequent behavior was smaller than .00000001. One analysis was designed to determine how many of the predicted results were significant at p equal to or less than .05 within each of eight different research areas. The results are shown in Table 7.8. The assumption was that, if the 345 had been a randomly selected sample of studies from a population of all possible studies for which the null hypothesis were true, we would expect 5% of the studies to achieve .05 significance by chance alone. The first column of numbers in Table 7.8 shows that all the proportions exceeded the expected value and that the median proportion of .39 is almost eight times larger than the expected value. Still, some unknown factors might have kept any negative results out of sight so that only these 345 studies were accessible. However, from a file-drawer analysis (the procedure is described in chapter 21), it was calculated that

TABLE 7.8
Expectancy effects in eight areas

Research area	Proportion of results that reached $p < .05$ in the predicted direction	Mean effect size in Cohen's d	Mean effect size in Pearson r
Lab interviews	.38	0.14	.07
Reaction time	.22	0.17	.08
Learning and ability	.29	0.54	.26
Person perception	.27	0.55	.27
Inkblot tests	.44	0.84	.39
Everyday situations	.40	0.88	.40
Psychophysical judgments	.43	1.05	.46
Animal learning	.73	1.73	.65
Median	.39	0.70	.33

it would take over 65,000 studies with null results to move the overall associated p to a barely acceptable .05. Other analyses concentrated on the size of the expectancy effect in each area, and those results are also listed in Table 7.8 as Cohen's d (Equation 2.4) and the Pearson r (described in detail in chapter 11).

Table 7.9 lists several strategies for controlling the effects of experimenters' expectancies and also notes one or more consequences of adopting these strategies (Rosenthal, 1979b; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979). First, assume

TABLE 7.9
Strategies for the reduction of experimenter expectancy effects

1. Increasing the number of experimenters
   - Decreases learning of influence techniques
   - Helps to maintain "blindness"
   - Randomizes expectancies
   - Increases generality of results
2. Monitoring the behavior of experimenters
   - Sometimes reduces expectancy effects
   - Permits correction for unprogrammed behavior
   - Facilitates greater standardization of experimenter behavior
3. Analyzing experiments for order effects
   - Permits inference about changes in experimenter behavior
   - Permits correction for expectancy effects
4. Maintaining "blind" contact
   - Minimizes expectancy effects
5. Minimizing experimenter-subject contact
   - Minimizes expectancy effects
6. Employing expectancy control groups
   - Permits assessment of expectancy effects

that the experimenter unwittingly learns from the participants' responses how to influence them unintentionally. This learning takes time, and with fewer participants from whom to learn the unintentional communication system, the experimenter may learn less of the system. Therefore, by increasing the number of experimenters so that each experimenter works with fewer subjects, it may be possible to reduce the likelihood of expectancy effects. Having more experimenters also helps to maintain blind contact between the experimenters and the subjects (i.e., the experimenters are unaware of which of the subjects are receiving the experimental and control treatments). The fewer the participants contacted by an experimenter, the less the chance of an unwitting breakdown in the blind procedure. A further advantage of increasing the number of experimenters is that the positive and negative expectancies may act like random errors that cancel one another. And finally, even beyond expectancy bias, we can be more confident of a result obtained by a larger number of experimenters than of a result obtained by only one experimenter.

Second, monitoring the behavior of experimenters may not by itself eliminate expectancy biases, but it may help in identifying unprogrammed expectancy behaviors. If we make our observations during a preexperimental phase, we may be able to use this information to select good experimenters. The problem is that this selection procedure may be unintentionally biased, and therefore, it may be preferable simply to assign experimenters to experiments randomly. Nevertheless, monitoring may alleviate some of the other biasing effects of experimenters noted previously, and it should facilitate greater standardization among the experimenters.

Third, analyzing experiments for order effects enables us to compare early results with later results. We can do a median split of the participants seen by each experimenter and compare the behavior of the participants in each half. Are the means of the groups the same? Is the amount of variability in the performance of the participants the same in both halves? We may also be able to correct for expectancy effects. In some cases, for example, we will find expectancies distributed only dichotomously; either a result is expected or it is not. At other times, we will have an ordering of expectancies in terms of ranks or absolute values. In any of these cases, we can correlate the results obtained by the experimenters with their expectancies. If the correlation is trivial in size, we are reassured that expectancy effects were probably not operating. If the correlation is substantial, we conclude that expectancy effects did occur. These can be "corrected for" or at least analyzed by such statistical methods as partial correlation (chapter 11) or blocking strategies (chapter 16).

The fourth strategy is based on the idea that, if the experimenter does not know whether the subject is in the experimental or the control group, the experimenter can have no validly based expectancy about how the person *should* respond. In drug trials, for example, in a **single-blind study** the participants do not know the group or condition (e.g., drug vs. placebo) to which they have been randomly assigned. In a **double-blind study,** both the experimenters and the subjects are kept from knowing what drug has been administered. Psychologists have been slow to adopt the double-blind method for other than drug trials, but when it is feasible, it is more than warranted to minimize the possibility of expectancy effects. A problem, however, is that single-blind and double-blind methods are not very easy to implement. Imagine a study in which anxiety

TABLE 7.10
Basic expectancy control design

	Expectancy conditions	
Treatment conditions	**Experimental treatment**	**Control treatment**
Experimental	Group A	Group B
Control	Group C	Group D

is the independent variable. People who have just been through an anxiety-arousing event, or who have scored high on a test of anxiety, may behave in an identifiable way in an experiment. The "blind" experimenters may then covertly "diagnose" the level of anxiety. If they know the hypothesis, they may unwittingly bias the results of the experiment in the expected direction or, by bending over backward to avoid bias, "spoil" the study. A score of subtle signs (the subject's arrival time, fidgeting behavior, and so on) may break down the most carefully arranged double-blind study.

A fifth strategy is to minimize the experimenter-subject contact, perhaps easier than trying to maintain blind contact. The day may come when the elimination of the experimenter, in person, will be a widespread, well-accepted practice. By computer, we can generate hypotheses, sample hypotheses, sample the experimental treatment conditions from a population of potential manipulations, select our participants randomly, invite their participation, schedule them, instruct them, record and analyze their responses, and even partially interpret and report the results. In experiments that require human interaction, it may still be possible to minimize the contact. For example, we might use an ordinary tape recorder and have a screen interposed between the experimenter and the participants.

The final strategy is the use of **expectancy control groups.** Although expensive to implement if many experimenters are randomly assigned to conditions, the advantage of this method is that we can compare the effects of experimenter expectancies with the effects of some other behavioral variable. Table 7.10 shows the most basic expectancy control design, in which there are two row levels of the behavioral research variable and two column levels of the experimenter expectancy variable. Group A is the condition in which the experimental treatment is administered to the subjects by a data collector who expects the occurrence of the treatment effect. In Group D, the absence of the experimental treatment is associated with a data collector who expects the nonoccurrence of the treatment effect. Group B is the condition in which subjects receiving the experimental treatment are contacted by an experimenter who does not expect a treatment effect. Subjects in Group C do not receive the experimental treatment and are contacted by an experimenter who expects a treatment effect.

Table 7.11 shows the results of a study by J. R. Burnham (1966) that used the expectancy design in Table 7.10. Burnham had 23 experimenters each run one rat in a T-maze discrimination problem. About half the rats had been lesioned by the removal of portions of the brain; the remaining animals had received only sham surgery, which involved cutting through the skull but no damage to the brain tissue. The purpose of the study was explained to the experimenters as an attempt to learn the effects of

TABLE 7.11
Expectancy control design used by Burnham (1966) to study discrimination learning in rats

	Expectancy conditions		
Treatment conditions	**"Lesioned"**	**"Nonlesioned"**	**Sum**
Lesioning of brain	46.5	49.0	95.5
No lesioning of brain	48.2	58.3	106.5
Sum	94.7	107.3	

Note: Cell values in this table are transformations of ranks to normal deviates, using a procedure described by Walker and Lev (1953, pp. 424–425), on the assumption that the underlying metric is normally distributed. The reason the cell values do not resemble *Z* scores (discussed in chapter 10) is that the transformation of ranks is based on a mean of 50 and standard deviation of 10. The range will vary, depending on the number of ranked scores. In Burnham's study, the sample size was 23; the top-ranked rat having a standard score of 70 and the bottom ranked rat, a standard score of 30. Thus, higher cell values in this table imply better performance in the T-maze discrimination problem.

lesions on discrimination learning. The design manipulated the expectancies by labeling each rat as lesioned or nonlesioned. Some of the really lesioned rats were labeled accurately as "lesioned" (the upper-left cell), and some were falsely labeled as "nonlesioned" (the upper-right cell). Some nonlesioned rats were labeled accurately (the lower-right cell), and some were falsely labeled as "lesioned" (the lower-left cell). Table 7.11 shows the standard scores of the ranks of performance in each of the four conditions (higher scores denote superior performance). Animals that had been lesioned did not perform as well as those that had not been lesioned, and animals that were believed to be lesioned did not perform as well as those that were thought to be nonlesioned. What makes this experiment of special interest is that the effects of expectancy were similar to those of the actual removal of brain tissue. Thus, it emphasizes the value of separating expectancy effects from the effects of the independent variable of interest, to avoid misrepresenting the causal impact of either variable.

CONCLUDING COMMENTARY

We do not want to end this chapter by leaving readers with a princess-and-the-pea image of human subjects as overly sensitive and overly responsive to the slightest experimental variations. It is possible for even the most outrageous manipulation to have no effect, and it is not easy to foresee when biasing effects will actually emerge (Sommer, 1968). In 1928, H. B. Hovey described administering an intelligence test to 171 people divided into two groups. One group took the test in a quiet room, and the other group took it in a second room with seven bells, five buzzers, a 550-watt spotlight, a 90,000-volt rotary-spark gap, a phonograph, two organ pipes of varying pitch, three metal whistles, a 55-pound circular saw mounted on a wooden frame, a photographer taking pictures, and four students doing acrobatics! Events in the second room were choreographed so that a number of distractions sometimes occurred concurrently and at other times the room was quiet. The remarkable result reported by

Hovey was that the group in the second room scored as well as the group in the first. Although we do not know whether anyone ever replicated Hovey's finding, we assume that it was accurately reported. Nonetheless, one major purpose of this chapter was to sensitize researchers to the kinds of threats to validity of the causal inferences discussed here. When we act as though we are oblivious to those threats, our science and the society that supports it both suffer.

As another poignant illustration, the physicist Richard Feynman (1999) described an incident in a psychology department in which an experimenter was running rats through mazes consisting of a long corridor with doors along one side where the rat entered, and doors along the other side in which the food was placed. The experimenter was trying to condition rats to enter the third door down from wherever they started, but try as he might, the rats invariably went immediately to the door where the food had been on the previous trial. The experimenter suspected that an uncontrolled variable of some kind was cueing the rats, so he painted the doors, making sure they appeared exactly alike. That did not work, so he then used chemicals to change the smell after each trial, and when that still did not work, he tried altering the lighting and the arrangement in the laboratory. It was maddening, until he finally figured out that the rats could tell which door they had previously entered by the way the floor sounded to them. The way he was finally able to fool them was to cover the corridor in sand, so the rats had to go in the third door if they wanted the food. Feynman told how, years later, he looked into the history of this research and learned that the control criteria developed by that experimenter were never absorbed by colleagues of the experimenter. They just went right on running rats in the same old way, oblivious to the methodological insights because the experimenter had not seemed to discover anything about rats. However, as Feynman (p. 215) noted, the experimenter had discovered things you have to do to find out something about rats.

In the 1970s, many psychological researchers expressed being overwhelmed by all the plausible sources of subject- and experimenter-related artifacts. We once compared this situation with a juggler's trying to balance dozens of spinning plates on the ends of sticks. The juggler has to keep running back and forth to keep them all balanced, just as the researchers in the 1970s felt they had to concentrate on one source of artifacts after another in order to keep everything properly balanced. What was needed, it seemed, was a conceptual pulling together of what was known about demand cues and artifacts within the framework of a workable, comprehensive model. Such a model has evolved in a collaboration by Rosnow successively with Leona Aiken, Daniel J. Davis, and David Strohmetz (Rosnow & Aiken, 1973; Rosnow & Davis, 1977; Strohmetz & Rosnow, 1994). Instead of focusing on specific artifact-producing variables, this "mediational model" concentrates on intervening (or mediational) steps in a theorized causal chain from the sources of uncontrolled task-orienting cues to their resulting artifacts. Readers interested in learning about the model will find a general description in Rosnow and Rosenthal (1997, ch. 4) and an operationalization and elegant series of studies in C. T. Allen (2004).

Another intriguing aspect of the artifact work was mentioned by McGuire (1969), who described the three stages in the life of an artifact as ignorance, coping, and exploitation. At first, most researchers seem unaware of the artifact and deny

its existence even when it is pointed out. Next, they view it as a nuisance variable and look for ways to isolate, eliminate, or control it. Finally, they realize that it can also be exploited as an independent variable of substantive interest. For example, we mentioned how the role of demand characteristics has evolved, so that what were once regarded as mere nuisance variables are now perceived as powerful substantive agents with practical implications in their own right. Demand characteristics are now conceived of as a potent source of behavioral change and accommodation in a wide variety of circumstances, including not only the experimenter-subject interaction in psychological research but also therapeutic change in the clinical situation (Orne & Bauer-Manley, 1991; Orne & Whitehouse, 2000). As Orne wisely noted many years ago, to understand the meaning of any social interaction, it is vital to take into consideration the role of demand characteristics in each and every situation. The same lesson applies to empirical research on the role of interpersonal expectations, which took root in the work on experimenter expectancy effects and has stimulated the burgeoning growth of interpersonal insights. In fact, the awareness of sources of artifacts has enhanced our understanding not only of the experimental setting but also of the nature of behavior and of the limitations of understanding.

Finally, we want to return to a point made in chapter 2: Most researchers would agree that it is simply impossible to design an experiment that will forever be free of plausible rival explanations. Probative experiments are designed to test hypotheses, theories, and models anchored in the experimenter's experiential world; their conceptual limits can never be exactly known because it is only by the discovery of experiences outside their jurisdiction that their boundaries are revealed. In spite of this uncertainty, our hypotheses, theories, and models form a constituency of intellectual assumptions about the world in which we live. Furthermore, our hypotheses, theories, and models are idealizations of reality, which restrict or stylize reality by forgoing all those features that cannot be entirely captured by the formulation. If there is no such thing as an experiment that can be confidently regarded as entirely free of alternative explanations, then the falsificationist view is an oversimplification of the way that scientific knowledge evolves. A paper by Brinberg, Lynch, and Sawyer (1992) makes the further point that "both findings consistent and findings inconsistent with a theory's predictions can be informative" (p. 140), and using a Bayesian analysis of hypothesis testing, they showed that both a priori and post hoc explanations may have equal merit in certain circumstances. Though internal validity may be viewed by most behavioral and social researchers as the sine qua non of valid causal inference, the reality may be that it is an ideal that is forever beyond our grasp. There is always the possibility that some theory or observation awaiting discovery will threaten the internal validity of even the most brilliantly designed experiment.