CHAPTER 9

CONDITIONING AND LEARNING

TYPES OF CONDITIONING
- Classical Conditioning: Does the Name Pavlov Ring a Bell?
- Instrumental (Operant) Conditioning

INTRODUCING THE VARIABLES

9.1 EXPERIMENTAL TOPICS AND RESEARCH ILLUSTRATIONS
- Within- and Between-Subjects Designs: Stimulus Intensity

9.2 EXPERIMENTAL TOPICS AND RESEARCH ILLUSTRATIONS
- Counterbalancing: Simultaneous Contrast

9.3 EXPERIMENTAL TOPICS AND RESEARCH APPLICATIONS
- Small- \( n \) Designs: Behavior Problems in Children

FROM PROBLEM TO EXPERIMENT: THE NUTS AND BOLTS
- The Partial Reinforcement Extinction Effect

SUMMARY
KEY TERMS
DISCUSSION QUESTIONS
WEB CONNECTIONS
PSYCHOLOGY IN ACTION: KNOWLEDGE OF RESULTS AS REINFORCEMENT
The above quote by Mark Twain describes a learning experience that shares many characteristics with a kind of conditioning called classical conditioning. In classical conditioning, the experience, as Twain indicates, is usually not under control of the organism.

The cat does not seek out a hot stove lid to sit on; rather, the stove just happens to be hot when she sits on it; there is nothing she can do about it. Twain also notes that the effects of conditioning can be far reaching—the cat will not choose to sit on any stove lid again, regardless of whether the lid is hot. This sort of conditioning is a prominent part of our lives. Some of the more important examples of classical conditioning have been studied by Ader and his associates.

Following work by Herrnstein (1962), Ader and Cohen (1982) wanted to see if a conditioned placebo (see Chapter 6) could influence the body's immune response. Herrnstein showed that when a saline solution was paired with a drug that can slow down behavior (scopolamine), the saline solution alone resulted in a slowing down of behavior. Just as Twain’s cat associated heat with stove lids—even cold stove lids—associating the salty solution with scopolamine led to its producing behavior similar to the behavior produced by scopolamine. Ader and Cohen paired the taste of a novel liquid with a drug that suppresses the immune system. After several pairings, the novel stimulus became a placebo, because it could also suppress the immune system. In turn, the immune suppression by the placebo prevented the development of lethal kidney problems in mice that were susceptible to kidney disease. Such findings helped lead to the development of psychoneuroimmunology, which is an interdisciplinary study of the interrelationships among behavior, neural, endocrine, and immune processes (Irwin & Miller, 2007). A major focus of this new field is to understand how conditioning can modulate immune responses.

Not all conditioning involves events that simply happen to the organism. In instrumental conditioning, the behavior of the organism is instrumental in producing consequences, often called rewards and punishments, which alter the rate of the behaviors that have produced the consequences. Let us examine an example of instrumental conditioning in an infant.

The behavior of infants between 2 and 6 months old is not very complicated, but they do show excited body movements to interesting stimuli. Rovee-Collier (1993) took advantage of this excitement to look at learning and retention in infants. She taught infants an instrumental kick response, by following a kick with the movement of a mobile that was suspended over the infant in his or her crib. A ribbon attached to the infant’s leg and the mobile permits the infant’s kick to activate the mobile. The rate of kicking when attached to the mobile is compared with the rate of kicking in a baseline phase (see Chapter 3) when the ribbon does not connect the infant and the mobile. Kicking is the instrumental response, and Rovee-Collier observed increases in kicking.
above the baseline in 2-, 3-, and 6-month-old infants. The infants learned an instrumen-
tal response—kicking—to activate the mobile.

Classical conditioning allows an organism to represent its world by learning rela-
tions among events (Rescorla, 1988)—a novel taste predicts an immune response. Instrumental conditioning, on the other hand, leads primarily to an organism learning relations between behavior and the outcomes it produces—a kick activates a mobile. These two fundamental types of learning will provide the background to illustrate several methodological issues. Before discussing these issues, we describe the basic features of these two types of conditioning.

▼ TYPES OF CONDITIONING

Classical Conditioning: Does the Name Pavlov Ring a Bell?

Early in this century, some fundamental psychological discoveries were made by Ivan P. Pavlov (1849–1936). The basic discovery that he made, which is now called Pavlovian, respondent, or classical conditioning, is well-known today, even by those outside psychology—but the fascinating story of how it came about is not. Pavlov was not trained as a psychologist (there were no psychologists in Russia, or elsewhere, when he received his education), but as a physiologist. He made great contributions in physiology concerning the measurement and analysis of stomach secretions accom-
panying the digestive process. He carefully measured the fluids produced by different sorts of food, and he regarded the secretion of these stomach juices as a physiologi-
cal reflex. For this important work on the gastric juices involved in digestion, Pavlov
received the Nobel Prize for medicine in 1904.

After he won the Nobel Prize, Pavlov turned to some more or less incidental discoveries that he had made in the course of his work. His fame in experimental psychology today arises from the systematic work that he did on these incidental discoveries, especially those concerning salivation. In one type of physiological experiment, Pavlov cut a dog’s esophagus so that it would no longer carry food to the stomach (Figure 9.1). But he discovered that when he placed food in the dog’s mouth, the stomach secreted almost as much gastric juice as it did when the food went to the stomach. Of course, food in the mouth produced reflexive salivation, but the reflexive action of the stomach seemed to depend on stimuli located in other places besides those in direct contact with the stomach lining. How-
ever, then Pavlov made an even more remarkable discovery. He found that it was not even necessary to place food in contact with the mouth to obtain salivary and gastric secretions. The mere sight of food would produce the secretions, or even the sight of the food dish without the food—or even the sight of the person who usually fed the animal! Obviously, then, secretion of the gastric juices and saliva must be caused by more than an automatic physiological reflex produced when a substance comes into direct contact with the animal. A physiological reflex is one that is shown by all physically normal animals of a certain species; it is “wired” into the nervous system. Pavlov had discovered a new type of reflex, one that he sometimes called a psychic reflex and sometimes a conditioned reflex. An outline of the standard paradigm for studying Pavlovian conditioning in the laboratory appears in Figure 9.2.
**FIGURE 9.1**

Ivan Pavlov and his staff are shown here with one of the dogs used in his experiments. The dog was harnessed to the wooden frame shown in the picture. Saliva was conducted by a tube to a measuring device that could record the rate and quantity of salivation.

**FIGURE 9.2**

An Outline of the Stages in Classical Conditioning. A neutral stimulus, such as a bell, that elicits no salivation when presented by itself is delivered to the organism slightly before an unconditioned stimulus, such as food powder, that produces an unconditioned response, salivation. If the neutral stimulus predicts the unconditioned stimulus, then after a number of pairings (as in phase 1), the neutral stimulus becomes associated with the unconditioned response, indicated by the broken line in phase 2. The neutral stimulus is now called the conditioned stimulus. Eventually (phase 3), the conditioned stimulus will elicit salivation in the absence of the unconditioned stimulus, and this salivation is called the *conditioned response* (CR). If the conditioned stimulus is repeatedly presented without the unconditioned stimulus, the CR will grow weaker and eventually extinguish.
Pavlov's very important discovery demonstrates that nearly any stimulus that does not normally elicit a particular response can come to control that response by being paired with another stimulus that reliably produces a reaction. For example, if a person is subjected to a stressful event in a particular situation (say, his or her office or room), then the body will respond with defensive reactions: The heart may race, blood pressure increases, adrenaline flows, and so on. These responses might become conditioned to the situation, so that even without a stressful event a person may show all these physiological changes in the situation. Some researchers in the field of behavioral medicine have argued that many disorders so prevalent today—hypertension, gastric ulcers, headaches—are in part caused by conditioned reactions that occur when people are in situations where they have repeatedly experienced stress. As indicated by the case in which a novel taste produced immune suppression, Pavlovian conditioning also seems to play a role in the development of other important human behaviors, including phobias (irrational fears) and other aversions.

Pavlovian conditioning does not, however, indicate that you are entirely at the mercy of arbitrarily paired events. Research has shown that although contiguous pairing of events (say, an office and stress) may be necessary for classical conditioning to occur, such pairing may not be sufficient. Several years ago, Kamin (1969) showed that it is important for the conditioned stimulus (CS) to predict the unconditioned stimulus (US) and not just coincide with it. Kamin first demonstrated that rats could easily learn to associate CSs that were a light, a noise, or a combination of light and noise with a mild electric shock (the US). He then showed that if a rat first learned that one CS was associated with shock (say, the noise), the rat would not learn much at all about the light–US relation when the light was subsequently presented with the noise CS in combination with the shock. Kamin reasoned that the rat first learned that noise predicted shock. Then when the noise and light appeared together, the light was redundant to the noise and the animal did not learn to associate it with the shock. Such blocking of learning to redundant stimuli compounded with the original CS also appears in experiments with humans (Mitchell & Lovibond, 2002). According to these results, a person who showed a stress reaction to his or her office would not learn to associate stress with a new piece of office furniture because the office itself would be a good predictor of stress. When one event predicts another, a contingent relation (or a contingency) is said to exist between the two events.

**Instrumental (Operant) Conditioning**

The earliest examples of the second type of conditioning were experiments by E. L. Thorndike at Columbia University, who put cats in puzzle boxes from which they were supposed to escape. These experiments were performed at about the same time as Pavlov’s. They are described in detail in Chapter 11; briefly, the experiment concerned learning from the consequences of some action. Thorndike’s cats performed some response that allowed them to escape from the puzzle box; then, when they were placed in the box again, they tended to perform the same response. The consequences of a behavior affected how it was learned. Since the behavior was instrumental in producing the consequence (the reward), this form of learning was called instrumental conditioning. It was seen as obeying different principles from those of Pavlovian conditioning and was also viewed as a more general type of learning.
Over the years, a great number of psychologists have devoted much effort to understanding instrumental conditioning. The most famous investigator and popularizer of the study of this type of conditioning is B. F. Skinner. He called this type of conditioning **operant conditioning**, because the response operates on the environment. This is distinguished from what Skinner called **respondent conditioning**, the classical conditioning studied by Pavlovians, in which the organism simply responds to environmental stimulation. The primary datum of interest in the study of operant conditioning is the rate at which some response occurs. The primary responses that have been studied are lever pressing by rats and key pecking by pigeons in operant-conditioning apparatuses (or, more colloquially, in Skinner boxes). A Skinner box is simply a small, well-lit box with a lever or key that can be depressed and a place for dispensing food (Figure 9.3).

In operant conditioning, the experimenter waits until the animal makes the desired response; he or she then rewards it, say, with food. If the entire response is not performed, the experimenter must reinforce successive approximations to it until the desired response occurs. For example, if you want to teach a pigeon to walk around in figure eights, you must first reward quarter-circle turns, then half-turns, and so on. This procedure of reinforcing greater and greater approximations to the desired behavior is called **shaping** the behavior; in principle, it is similar to the “getting warmer” game played by children. Operant conditioning works on the law of effect (see Chapter 11 for additional details): If an operant response is made and followed by a reinforcing stimulus, the probability that the response will occur again is increased. What will serve as a reinforcing stimulus is not specified ahead of time but must be discovered for each situation. Using this straightforward principle, Skinner and many others have undertaken the experimental analysis of numerous behaviors.

![Figure 9.3](image-url)

A typical Skinner box equipped with a response lever and a food cup below it. A lever press by the animal makes a pellet of food drop into the cup. All of this machinery is controlled by programming equipment that allows the experimenter to set different tasks for the animal.
A reinforcing stimulus is one that strengthens the response that it follows. Generally, two different classes of reinforcing stimuli are identified: positive reinforcing stimulus and negative reinforcing stimulus. Positive reinforcers are the familiar rewards that are given following a particular response: food pellets in the Skinner box, a gold star on a good spelling test, a moving mobile in a crib, and so on. Behaviors that produce positive reinforcers increase in likelihood. Negative reinforcers are aversive events, and responses that remove or avoid them are strengthened. For example, a rat can learn to press a bar in a Skinner box to terminate a shock or to postpone it. In a similar fashion, you receive reinforcement on a bitterly cold day by putting on a warm coat to avoid getting uncomfortably chilly.

Do not confuse negative reinforcement with punishment. Behaviors that produce aversive events are said to be punished. Punished behaviors decrease in frequency. If a rat has been positively reinforced with food for pressing a lever and is now punished for each lever press with a mild electric shock, the rate of pressing the lever will decrease.

The role of stimuli other than the reinforcing stimulus is also important in the study of operant conditioning. A discriminative stimulus (SD) signals when a behavior will be followed by a reward. For example, a pigeon might be trained to peck a button for food only in the presence of a red light. If any other light is on, pecking will not be followed by food. Animals learn such contingencies between stimuli and responses quite readily. A discriminative stimulus may be said to “set the stage” or “provide the occasion” for some response. One of the primary tasks of operant conditioning is to bring some response “under stimulus control.” An organism is said to be under stimulus control when it responds correctly and consistently in the presence of a discriminative stimulus and not in its absence. Rovee-Collier’s infants were under stimulus control when they responded to the mobile that resulted in reinforcement.

Recall that effective CSs in classical conditioning are those that are contingently related to the US (the CS predicts the US). Operant conditioning also has important contingent relations: namely, the relation between the response and the reinforcing stimulus. In operant conditioning, organisms can learn that there is a positive contingency between behavior and reinforcement, which is the standard Skinner box arrangement. Organisms can also learn that there is a negative contingency between behavior and reinforcement, such that if they respond, reinforcement will not occur. One way of examining this contingency is to use a procedure called experimental extinction, in which reinforcement is withheld after an organism has learned to make a particular response. After several times of having reinforcement fail to follow the response, the organism ceases to make the response. (See the “From Problem to Experiment” section at the end of this chapter for an additional discussion of extinction.)

Finally, organisms can learn that there is a null contingency between their behavior and reinforcement. A null contingency is one in which the reinforcer is independent of behavior—sometimes a behavior leads to reinforcement, and sometimes the reinforcement occurs in the absence of a particular behavior. When organisms learn that no matter what they do, aversive events will happen, they learn to be helpless and show various signs of depression (LoLordo, 2001). Likewise, some research shows that when positive events occur independently of behavior, organisms tend to become lazy (Welker, 1976). It is tempting to generalize the latter effect to “spoiled brats” who receive substantial numbers of noncontingent positive reinforcers and then fail to work hard to get them when it becomes necessary to do so.
Dependent Variables

One important dependent variable in animal learning research is the rate of responding, often plotted over time. Another important dependent variable commonly used in classical conditioning is the amplitude of the response. Rather than just noting whether or not a dog salivates to a conditioned stimulus, we can measure the amplitude of the response by seeing how much saliva is produced. Another commonly measured characteristic of responses is their latency, or the time it takes the animal to accomplish the response. This measure is widely used in maze-learning experiments, where the time it takes an animal to complete the maze is recorded. Often results are plotted in terms of speed rather than latency, speed being the reciprocal of latency (1/latency).

A derived measure of learning is resistance to extinction. After a response has been learned, if the experimenter no longer applies reinforcement when the animal executes the response, the response gradually grows weaker or extinguishes. Resistance to extinction, then, can be used as an index of how well the response was learned in the first place. It is considered a derived, rather than basic, measure because what is still being measured is frequency, amplitude, or speed of response. These all decline during extinction, but they may decline at different rates after different manipulations of the independent variable. Thus, resistance to extinction is a derived measure of the effectiveness of some independent variable on learning.

Independent Variables

A great many independent variables may be manipulated in studies of animal learning and conditioning. Many have to do with the nature of reinforcement. Experimenters can vary the magnitude of reinforcement in Pavlovian and instrumental conditioning.

Control Variables

Control of extraneous variation is typically quite sophisticated in basic animal learning research, but even here, there are subtle problems. One of these is the problem of pseudoconditioning in classical conditioning experiments. Pseudoconditioning refers to a temporary elevation in the amplitude of the conditioned response (CR) that is not caused by the association between the CS and US. Thus, it is not true conditioning but only mimics conditioning. It is recognized by being relatively short lived and variable and is usually caused by the general excitement of the experimental situation for the animal, including the presentation of the CS and US. The appropriate control for pseudoconditioning is to have one group of animals in the experiment exposed to the same number of CS and US presentations as the animals in the conditioning group but to have the presentations unpaired and presented randomly. Both the experimental and the pseudoconditioning control groups should be affected by the general excitement induced by the experimental situation; any difference between the two groups should be due to the learning produced by the CS-US pairings in the case of the experimental group (Rescorla, 1967).
CHAPTER 9 CONDITIONING AND LEARNING

9.1 EXPERIMENTAL TOPICS AND RESEARCH ILLUSTRATIONS

**Topic** Within- and Between-Subjects Designs

**Illustration** Stimulus Intensity

One fundamental question about classical conditioning is how the intensity of a neutral stimulus affects the conditioning process. For example, if a tone is paired with food given to dogs, will the intensity of the tone affect how quickly the dog becomes conditioned, so that the tone alone (now called the conditioned stimulus) produces salivation? A reasonable hypothesis is that the stronger the stimulus, the more quickly conditioning will occur. Animals should be very sensitive to more-salient stimuli and thus more likely to associate them with unconditioned stimuli. We might predict that the stronger the conditioned stimulus, the faster and stronger the conditioning will occur.

Many researchers have investigated this question over the years; the surprising finding from most of the early research was that stimulus intensity did not seem to have much effect on Pavlovian conditioning. Relatively weak stimuli seemed to produce just as good conditioning as did strong stimuli (Carter, 1941; Grant and Schneider, 1948). Since most theories of the time predicted that conditioning should be affected by the intensity of the stimulus (Hull, 1943), the failure to find the effect constituted something of a puzzle. The researchers who did these experiments on stimulus-intensity effects in conditioning typically used between-subjects designs, so that different groups of subjects received different stimulus intensities. Before discussing the reason for their doing this, let us consider some of the general advantages and disadvantages of between-subjects and within-subjects experimental designs.

**Between-Subjects Versus Within-Subjects Designs**

Consider the simplest sort of experimental design in which there are two conditions, experimental and control, with different groups of participants assigned to each. In a **between-subjects design**, a different group of participants usually receives just one level of each independent variable. One potential problem that can arise from using a between-subjects design is that a difference obtained on the dependent variable might be caused by the fact that different groups of participants are used in the two conditions. This means that in the standard between-subjects design participants are confounded with the levels of independent variable. Experimenters try to overcome this problem by randomly assigning subjects to the levels of the independent variable in between-subjects designs. Thus, on the average, the groups should have similar characteristics in all conditions. In all between-subjects designs, then, participants should be randomly assigned to the different conditions to ensure that the groups are equivalent prior to the manipulation of the independent variable. We must try to have equivalent groups, because otherwise any difference observed between the groups on the dependent variable might be merely because participants in the different groups differed in ability. If we have a large number of subjects and randomly assign them to groups, then we can minimize the possibility of this sort of confounding and be more confident that any difference we find on the dependent variable actually results from the independent-variable manipulation. When subjects are randomly assigned to conditions in a between-subjects design, this is referred to as a **random-groups design**.

There are two primary drawbacks to between-subjects or random-groups designs. One is the fact that they are wasteful in terms of the number of subjects required.
When a different group of subjects is assigned to each condition, the total number of participants required for an experiment can quickly become quite large—especially if the experimental design is at all complex. Thus, between-subjects designs are impractical when, as is often the case, there is a shortage of participants available for an experiment.

The second problem is more serious; it has to do with the variability introduced by using different groups of participants. One basic fact of all psychological research is that subjects differ greatly in their abilities to perform almost any task (on almost any dependent variable). When numerous participants are used in between-subjects designs, some of the differences in behavior in the experimental conditions will result from differences among the participants. This can have the unfortunate effect of making it difficult to determine whether subject differences or the independent variable determined the results. In summary, participants are confounded with groups, and in between-subjects designs this variability caused by their subject differences cannot be estimated statistically and taken into account.

Both of the problems with between-subjects designs can be reduced by using within-subjects designs, in which all participants receive every level of the variable. Within-subjects designs usually require fewer subjects than between-subjects designs, because each participant serves in all conditions. Also, statistical techniques can take into account the variance produced by the differences between subjects. This is possible because each one serves as his or her own control—another way of saying that participants are not confounded with groups, as in a between-subjects experiment. A within-subjects design is usually more effective than a between-subjects design in detecting differences between conditions on the dependent variable, because this variance owing to participant differences can be estimated statistically and taken into account. The exact statistical techniques for analyzing within-subjects experiments are not discussed here. In general, the within-subjects design is more powerful—more likely to allow detection of a difference between conditions if there really is one—than the between-subjects design. This advantage makes the within-subjects design preferred by many investigators whenever it is possible to use it.

Although there are advantages to using within-subjects designs, new problems are introduced by them. Unfortunately, within-subjects designs simply cannot be used in investigating some types of experimental problems, and even when they can be used, they have requirements that between-subjects designs do not. Within-subjects designs cannot be used in cases where performing in one condition is likely to completely change performance in another condition. This problem is usually called asymmetrical transfer, or a carryover effect. If we want to know how rats differ in learning a maze with and without their hippocampuses (a part of the brain related to learning and memory), we cannot use a within-subjects design, since we cannot replace a hippocampus once it is removed. The same problem occurs any time the independent variable may provide a change in behavior that will carry over until the subject is tested under the other condition.

If we want to test people on a task either with or without some specific training, we cannot test them first with training and then with no training. And we cannot always test them in the reverse order (no training and then training), because then we have confounded conditions with practice on the task. For example, if we want to see if a specific memory-training program is effective, we cannot teach people the program (the training phase), test them, and then test them again with no training. Once a person has
had the training, we cannot take it away; it will carry over to the next part of the experi-
ment. We cannot test people in the other direction, either, with a memory test, a training
phase, and then a second memory test. The reason is that if people improved on the
second test, we would not know whether the improvement was a result of training or
merely practice at taking memory tests. In other words, training and practice would be
confounded. A between-subjects design is appropriate in this case. One group of people
would have their memories tested with no training; the other group would be tested
after receiving the memory training program.

Within-subjects designs are also inappropriate when participants may figure out
what is expected of them in the experiment and then try to cooperate with the experi-
menter to produce the desired results. This is more likely to happen with within-subjects
than between-subjects designs, because in the former case, the people participate in
each condition. This problem makes within-subjects designs all but nonexistent in cer-
tain types of social psychological research.

Even when these problems do not eliminate the possibility of using a within-
subjects design in some situations, there are additional problems to be considered. In
within-subjects designs, the subjects are always tested at two or more points in time;
thus, the experimenter must be on guard for factors related to time that would affect
the experimental results. The two primary factors that must be considered are practice
effects and fatigue effects, which tend to offset one another. Practice effects refer
to improved performance in the experimental task simply because of practice, and
not the manipulation of the independent variable (as in the memory experiment just
discussed). Fatigue effects refer to decreases in performance over the course of the
experiment, especially if the experimental task is long, difficult, or boring. The effects
of practice and fatigue may be taken into account and minimized by systematically
arranging the order in which the experimental conditions are presented to subjects.
This technique is referred to as counterbalancing of conditions and is discussed later
in the chapter.

As we have said, when appropriate, within-subjects designs are generally preferred
to between-subjects designs, despite the fact that they involve a number of additional
considerations. The primary advantage, once again, is the fact that within-subjects designs
are typically more powerful or more sensitive, because the possibility of error resulting
from subject variability is reduced relative to that in between-subjects designs.

A third design, the matched-groups design, tries to introduce some of the ad-
vantages of a within-subjects design to a between-subjects comparison. The matched-
groups design attempts to reduce participant variability among groups by matching
them in the different groups on other variables. Thus, in a human memory experiment,
people might be matched on the basis of IQ before they were randomly assigned to
conditions. (Each subgroup of people matched for IQ is randomly assigned to a par-
ticular group.) Matching on relevant variables can help reduce the variability caused
by the simple random assignment of participants to each group (the random-groups
design). Also, it is very important that assignment to conditions still be random within
matched sets of people; otherwise, there is the possibility of confounding and other
problems, especially regression artifacts (see Chapters 2 and 12). In many situations,
matching tends to involve much work, since participants must be measured separately
on the matching variable.

One matching technique used in animal research is the split-litter technique.
This involves taking animals from the same litter and then randomly assigning them to
groups. Since the animals in the different groups are genetically similar, this helps reduce variability resulting from subject differences that occurs in random-groups designs.

**Stimulus Intensity in Classical Conditioning**

Let us now return to the problem we considered earlier. How does the intensity of the conditioned stimulus affect acquisition of a conditioned response? Common sense, as well as some theories, led to the prediction that more intense stimuli should lead to faster conditioning than should less intense stimuli. However, as mentioned earlier, the first research on this topic failed to find such effects. For example, Grant and Schneider (1948) varied the intensity of a light as a conditioned stimulus in an eyelid-conditioning experiment. In such experiments, people are attached to an apparatus that delivers puffs of air to the eye and records responses (eye blinks). The unconditioned stimulus is the air puff, the unconditioned response (UR) is the blinking, and the conditioned stimulus is a light that precedes the air puff. Originally, the light does not cause a person to blink; but after it is repeatedly paired with the air puff, eventually the light by itself causes the blinking, which then is the conditioned response. Grant and Schneider asked simply whether more intense lights would cause faster conditioning. Would a bright light cause people to develop a conditioned response faster than a dim light? They tested different groups in the two cases, one with each intensity of light, and discovered that, contrary to expectations, conditioning was just as fast in the condition with the dim light as it was in the condition with the bright light. Other researchers obtained similar results when they examined the effects of stimulus intensity on conditioning in between-subjects designs, even when they used different stimuli (such as tones instead of lights).

The choice of a between- or within-subjects design is usually determined by the nature of the problem studied, the independent variables manipulated, the number of participants available to the researcher, and other considerations described in the preceding section. Rarely do researchers consider the possibility that the very outcome of their research could depend on the type of design they choose. However, this is exactly the case in the issue of the effects of stimulus intensity on conditioning, as was discovered after Grant and Schneider’s (1948) research.

Years later, Beck (1963) again asked the question of whether the intensity of the stimulus affected eyelid conditioning. She was also interested in other variables, including the intensity of the unconditioned stimulus (the air puff) and the anxiety level of her participants. For our purposes, we will consider only the effect of the intensity level of the conditioned stimulus, which was varied within subjects as one factor in a complex experimental design. Beck used two intensity levels and presented them in an irregular order across 100 conditioning trials in her experiment. She found a large and statistically significant effect of stimulus intensity on development of the conditioned response, contrary to what other researchers had found.

Grice and Hunter (1964) noticed Beck’s effect and wondered if she had found an effect where others had found none because she had used a within-subjects design, whereas most others had used between-subjects designs. To discover this, they tested three groups of people: In two groups, the variable of intensity of the conditioned stimulus was varied between subjects, and for the remaining group, it was varied within subjects. They used a soft tone (50 decibels) or a loud tone (100 decibels) as the conditioned stimuli in an eyelid-conditioning experiment. People in each group participated in 100 trials. On each trial, participants heard a buzzer that alerted them that a
trial was beginning. Two seconds later, they heard a tone (soft or loud) that lasted half a second, and then after another half-second, they received a puff of air to the eye. The tone was the conditioned stimulus, the air puff the unconditioned stimulus. The loud group received the 100-decibel tone on all 100 trials, whereas the soft group received the 50-decibel tone on all 100 trials. These two groups represent a between-subjects comparison of stimulus-intensity level in eyelid conditioning. People in the third group (the loud/soft group) received 50 trials with the loud tone (L) and 50 trials with the soft tone (S). The trials occurred in one of two irregular orders, such that one order was the mirror image of the other. In other words, if the order of the first 10 trials was L, S, S, L, S, L, S, L, the other order would be S, L, L, S, L, S, S, L, S. In the loud/soft group, half the people received each order.¹

The results are shown in Figure 9.4, where you can see the percentage of the last 60 trials on which participants showed a conditioned response—that is, blinking—to the

¹ You may have noticed that there is a confounding in the design of the experiment by Grice and Hunter. People received 100 trials with either a soft or a loud CS in the between-subjects case, but subjects received half as many of each CS in the within-subjects case. More trials should yield better conditioning to each CS for the between-subjects case, but as shown in Figure 9.4, it did not—loud led to better learning in the within-subjects case, and soft led to better learning in the between-subjects case.
tone before the air puff came. One line represents the between-subjects comparison, in which each subject had experience with only the loud stimulus or the soft stimulus. Notice that the percentage of trials on which participants responded did not vary as a function of stimulus intensity in the one-stimulus, between-subjects case. (The slight difference seen is not statistically reliable.) On the other hand, when stimulus intensity was varied within subjects in the loud/soft group, a large effect of stimulus intensity was found. People responded just slightly more than 20 percent of the time to the soft stimulus but almost 70 percent of the time to the loud stimulus, a significant difference.

The results of Grice and Hunter’s experiment show that the choice of a between-subjects or within-subjects design can have far-reaching effects. In this case, the actual outcome of experiments designed to examine the effects of stimulus intensity was determined by the choice of design. When people experienced both stimuli, they reacted to them differently; but when they experienced only one stimulus or the other, they showed no difference in responding. Grice (1966) reports other situations in which a similar pattern of results occurs. In many experimental situations, researchers cannot tell whether their findings would be changed by switching from a between-subjects to a within-subjects design (or vice versa), because it is impossible to ask the experimental question with the other design, for reasons discussed earlier. However, Grice and Hunter’s (1964) research reminds us that the choice of an experimental design can have ramifications beyond mundane considerations of the number of subjects used and the like. The actual outcome of the research may be affected.

Consider an alternative way of interpreting these results. Perhaps the between- and within-subjects designs result in different effects because the within-subjects design automatically produces a carryover effect by allowing the participants in the experiment to experience all values of the relevant stimuli. Compared with people in the between-subjects version, people in the within-subjects design may have been able to perceive the loud stimulus as louder and the soft stimulus as softer because they had the opportunity to experience both intensities on successive trials and could therefore compare one stimulus with another. People who perceived only one value of the stimulus could not make such a comparison. Likewise, Kawai and Imada (1996) have found that the greater aversiveness of a longer electrical shock than a shorter one is more likely to be noticed in a within-subjects design than in a between-subjects one. Later, in the “From Problem to Experiment” section on page 253, you will see that the choice of design can determine the course of extinction in instrumental conditioning for humans (Svartdal, 2000) and nonhuman animals (Papini, Thomas, & McVicar, 2002). In the next section we examine another sort of contrast that can occur in conditioning, and it also suggests that being able to compare events by virtue of the experimental design results in different kinds of behavior.

9.2 EXPERIMENTAL TOPICS AND RESEARCH ILLUSTRATIONS

<table>
<thead>
<tr>
<th>Topic</th>
<th>Counterbalancing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illustration</td>
<td>Simultaneous Contrast</td>
</tr>
</tbody>
</table>

Whenever a within-subjects design is used, one needs to decide with care the order in which the conditions should be presented to the subjects. The arrangement must be such that, on the average, the conditions are presented at the same stage of practice,
so that there can be no confounding between the experimental conditions and stage of practice. Counterbalancing is also necessary to minimize the effect of other variables besides time that might affect the experiment. Often, the experimenter must counterbalance across variables in which he or she has little interest, even in between-subjects designs, so that these extraneous variables do not affect the conditions of interest. An example of this problem should make it clearer.

In learning an instrumental response, the particular magnitude of the reward greatly influences performance. Typically, performance improves as the magnitude of reward increases. However, the particular magnitude of reward used does not have an invariable effect on performance but depends instead on the experience that the organism has had with other reinforcement conditions. One example of this effect is provided by an experiment done by Bower (1961) on simultaneous contrast, in which some subjects experienced two contrasting magnitudes of reward.

Bower's experiment consisted of three groups of 10 rats, each of which received four trials a day in a straight-alley maze for 32 days, for a total of 128 trials. The independent variable was the magnitude of reward used. One group of rats received eight food pellets in the goal box on their four trials. Since they received a constant eight pellets on each trial, this condition is referred to as Constant 8. Another group received only one pellet after each trial (Constant 1). These two groups can be considered as controls for the third (Contrast) group. Subjects in this group received two trials each day, in two different straight alleys. The two alleys were quite discriminable, one being black and one being white. In one alley, the rats always received a one-pellet reward; in the other alley, they always received eight pellets. Bower wanted to see how the exposure to both levels of reinforcement would affect running speed, as compared with exposure to only one level all the time. Would rats run at a different speed for a one-pellet (or eight-pellet) reward if they had experienced another level of reward, rather than having had constant training at one particular level?

Before examining the results, consider some design features that Bower had to face for the contrast rats. Since magnitude of reward varied within subjects in this condition, there were two problems to consider. First, he had to make sure that not all subjects received the greater or lesser reinforcement in either the black or white alley, because alley color then would be confounded with reward magnitude, and rats may simply have run faster in black alleys than white (or vice versa). This was easily accomplished by having half the animals receive eight pellets in the black alley and one pellet in the white alley and having the other half of the animals receive the reverse arrangement. For the control animals that received only one reward magnitude, half received the reward in a white alley and half in a black alley. All this may sound rather complicated, so we outlined the design scheme in Figure 9.5.

The second problem concerns the order in which the rats in the contrast group should be given the two conditions on the four trials each day. Obviously, they should not first be given the two large-reward trials followed by the two small-reward trials (or vice versa), since time of testing would be confounded with reward magnitude. Perhaps a random order could be used for the 128 trials. But random orders are not preferred in such cases, since there can be, even in random orders, long runs in which the same occurrence appears. So it would not be surprising to find cases where there were two trials in a row of the same type (large or small magnitude of reward), although across all subjects there would be no confounding with practice. A preferable way to handle this problem involves counterbalancing the conditions rather than
FIGURE 9.5
An outline of the design of Bower’s experiment showing some of the features used to minimize confounding. Alley color is not confounded with group or magnitude of reward (number of pellets). For the contrast group, additional counterbalancing is needed to balance the order of one- versus eight-pellet trials within the daily session (see text).

Table: Design of Bower’s experiment

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Rats</th>
<th>Alley Color</th>
<th>Number of Trials (per day)</th>
<th>Number of Pellets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant 8</td>
<td>5</td>
<td>Black</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>White</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Constant 1</td>
<td>5</td>
<td>Black</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>White</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Contrast</td>
<td>5</td>
<td>Black</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>White</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

varying them randomly. Counterbalancing, you will remember, refers to any technique used to systematically vary the order of conditions in an experiment to distribute the effects of time of testing (such as practice and fatigue), so that they are not confounded with conditions.

When two conditions are tested in blocks of four trials, there are six possible orders in which the conditions can occur within trials. In the present case, if $S$ stands for a small (one-pellet) reward and $L$ for a large (eight-pellet) one, then the six orders are SSLL, SLSL, LLSS, LSSL, SLLS, LSSS. Bower solved the counterbalancing problem by using each of these orders equally often. On a particular day of testing, he would pick an order for the trials for half the rats (e.g., LSSL) and then simply test the other half using the opposite order (SLLS). The next day he would pick another order for half the rats, while the others received the reverse order, and so on. This led to no confounding between order and conditions, and all the orders were used equally often, so that the experiment did not depend on just one order. We shall return to this point in a moment.

Bower’s results are presented in Figure 9.6, where the mean running speed for each of the four conditions is plotted across blocks of 2 days (eight trials). For the rats that received constant reward, those rewarded with eight pellets performed better after the first few days than those rewarded with one. It was not exactly big news, of course, that rats ran faster for more, rather than less, food. The real interest was in how fast rats in the contrast conditions ran for large and small rewards. Although there was no statistically reliable effect between speeds of the Constant 8 and Contrast 8 conditions in Figure 9.6, the Contrast 1 rats ran reliably more slowly for the reward than did the Constant 1 rats, at least toward the end of training. This is referred to as a negative contrast effect, since the contrast subjects ran more slowly for the small reward than did the rats that always received the small reward.
One interpretation of this phenomenon considers emotional states induced in the contrast rats owing to their experience in the situation. Since the contrast rats were familiar with both levels of reward, when placed in the distinctive alley that told them that they would receive a small reward, they were annoyed or frustrated at having to run down the alley for only one crummy pellet. These results prompt an interesting question as to why Bower did not find a positive contrast effect, or faster running for the Contrast 8 subjects relative to the Constant 8 subjects. Should not the Contrast 8 rats be happy or elated to learn, when placed in the distinctive alley signaling a large reward, that they would get eight pellets rather than only one? One possibility is that they were more elated but that a ceiling effect prevented this from being reflected in their running speeds. Perhaps performance was already very good in the Constant 8 condition. Because of the large reward, there was no room for improvement in the Contrast 8 condition. The rats in the control (Constant 8) condition were already running as fast as their little legs would carry them, so no matter how much more elated the contrast rats might be, this could not be reflected in their performance. Although this ceiling-effect interpretation of the present data is bolstered somewhat by other reports of positive contrast effects (Padilla, 1971), negative contrast effects seem more easily obtained than positive ones. We discuss the problem of ceiling effects in data analysis more fully in Chapter 10.
The results of Bower’s experiment should remind you of the lesson in the previous section contrasting between- and within-subjects designs. As in Grice and Hunter’s experiments on stimulus-intensity effects, Bower found that the effect of a reward of a particular magnitude depended on the type of design used. In a within-subjects design, in which the animals had experience with both reward magnitudes, the effect on behavior was greater than in the between-subjects comparison, in which a different group of animals received constant rewards over the series of trials. Once again, the nature of the design can affect the experimenter’s conclusion about how strong an effect is produced by the independent variable.

**Further Considerations in Counterbalancing**

A great variety of counterbalancing schemes can be used in various situations. Some of these become very complex. Here, we discuss only some of the simpler counterbalancing designs to provide you with a few tricks of the trade.

The case represented by Bower’s (1961) contrast group is in many ways typical of the counterbalancing problem as it usually arises. Two conditions were to be tested within subjects; thus, they had to be counterbalanced so as not to be confounded with stage of practice. One solution to this problem, and the one most psychologists would pick, would be to use an **ABBA design**, where \(A\) stands for one condition and \(B\) stands for the other. This would remove the confounding of particular conditions with time of testing, since each condition would be tested at the same time on average (\(1 + 4 = 5\) for \(A\), and \(2 + 3 = 5\) for \(B\), where the numbers refer to the order of test). But perhaps the specific order of testing might also matter. For example, let us assume that there is a very large practice effect on the dependent variable but that it occurs very early in training, on the first trial. Then it would contribute to the \(A\) condition but not to the \(B\) condition, so that the **ABBA** design would not eliminate the confounding of conditions with practice.

Two solutions to this problem of large effects of practice early in training can be suggested. One is to give a number of practice trials in the experimental situation before the experiment proper begins. Thus, the subjects are given practice, and performance on the dependent variable is allowed to stabilize before the experimental conditions of interest are introduced. Another solution is to employ more than one counterbalancing scheme. For example, half the subjects might get the reverse of the scheme that the other half receives. So half the subjects would get **ABBA** and the other half would get **BAAB**. Bower’s solution to the counterbalancing problem was the ideal extension of this logic, since he used every possible counterbalancing scheme equally often. But when more than two conditions are involved, this becomes unwieldy. In most situations, an adequate solution to the problem of practice effects at the beginning of a testing session would be to give subjects practice and then use two counterbalancing schemes, one of which is the reverse of the other. Grice and Hunter (1964), in the experiment on stimulus intensity described earlier, did just this.

For situations in which there are more than two conditions, we would recommend one particular scheme for counterbalancing as generally desirable. This is the **balanced Latin square design**. Suppose there were six conditions in a counterbalanced order so that practice effects would not confound the results. For example, in a simultaneous contrast experiment such as Bower’s, six different reward magnitudes could be used rather than only two. A balanced Latin square design would ensure
that when each condition was tested, it would be preceded and followed equally of-

ten by every other condition. This last feature is very useful in minimizing carryover

effects among conditions and makes the balanced Latin square preferred to other

counterbalancing schemes. Constructing a balanced Latin square is easy, especially

if the experiment has an even number of conditions. Let us number the six condi-
tions in an experiment from one to six. A balanced Latin square can be thought of

as a two-dimensional matrix in which the columns (extending vertically) represent

conditions tested, and the rows represent the subjects. A balanced Latin square for

six conditions is presented in Table 9.1. The subjects are labeled \( a \) through \( f \), and

the order in which they receive the conditions is indicated by reading across the

row. So subject \( a \) receives the conditions in the order 1, 2, 6, 3, 5, 4. The general

formula for constructing the first row of a balanced Latin square is 1, 2, \( n \), 3, 4, 6, 5,

and so on, where \( n \) stands for the total number of conditions. After the first

row is in place, just number down the columns with higher numbers, starting over

when you get to \( n \) (as in Table 9.1). When a balanced Latin square is used, subjects

must be tested in multiples of \( n \), in this case six, in order to counterbalance condi-
tions appropriately against practice.

When an experimental design has an odd number of conditions, it becomes a bit

more complicated to use a balanced Latin square. In fact, two squares must be used,

the second of which is the reverse of the first, as seen in Table 9.2, where once again

letters indicate subjects and numbers stand for conditions in the experiment. When

a balanced Latin square is used with an unequal number of conditions, each subject

must be tested in each condition twice. The case represented in Table 9.2 is for five

conditions. In general, the first square is constructed in exactly the same manner as

when there is an even number of conditions, and then the second square is an exact

reversal of the first.

The balanced Latin square is an optimal counterbalancing system for many pur-
poses, since each condition occurs, on the average, at the same stage of practice and
each condition precedes and follows every other equally often. This latter feature is not
true of other counterbalancing schemes, and thus there is more concern that testing in
one condition may affect testing in another condition.
PART 2  PRINCIPLES AND PRACTICES OF EXPERIMENTAL PSYCHOLOGY

In the simplest case, an experiment involves a comparison between control and experimental conditions. Both between- and within-subjects designs usually include a large number of subjects, although the number is often less in the within-subjects case. Large numbers of subjects are required so that an unusual participant does not skew the results. Such designs are called large-\( n \) designs and have become the norm in psychological research; powerful statistical techniques (see Appendix B) allow the researcher to determine whether differences between the conditions are worth worrying about. An alternative approach to such designs is the small-\( n \) design, in which a very few subjects are intensely analyzed. Two areas within experimental psychology often use small-\( n \) designs. You may recall from Chapter 6 that psychophysics often uses small-\( n \) designs. Experimental control is the hallmark of the second research area that uses small-\( n \) designs—the experimental analysis of behavior in terms of operant conditioning. Skinner (1959) urged the use of small-\( n \) designs in operant research, because he wanted to emphasize the importance of experimental control over behavior and deemphasize the importance of statistical analysis. Skinner believed that statistical analysis often becomes an end in itself, rather than a tool to help the researcher make decisions about the experimental results. The experimental control usually achieved in traditional research with large numbers of subjects and statistical inference is strived for in small-\( n \) research by very carefully controlling the experimental setting and by taking numerous and continuous measures of the dependent variable. Small-\( n \) methodology is especially appropriate to the clinical application of operant techniques to modify behavior. Typically, a therapist deals with a single client at a time, which is the limiting case of a small-\( n \) design. Although a therapist may treat more than one patient at a time with similar methods, the numbers are very small relative to those.
seen in most large-
 research. We examine small-
 design in the context of behavior
 problems in children.

Consider the following scenario: Concerned parents seek psychological help for
their child because she has temper tantrums several times a day. These tantrums are
noisy, with a lot of yelling and crying, and they are violent—she often kicks things and
bangs her head on the floor. When the therapist, who specializes in behavior modification
using operant-conditioning techniques, first sees the child, what conclusions might he or
she draw? What causes this behavior? What can the therapist do to remove it and return
the child to a more normal existence? A behavior therapist seeks to discover what in the
learning history of the child has produced such troublesome behavior, focusing on the
contingencies of reinforcement that produce and maintain the child’s crying, kicking, and
head banging. Perhaps the parents only paid close attention to the child when she threw
a tantrum—inadvertently rewarding her tantrum behavior. The proposed therapy tries
to change the contingencies of reinforcement, so that the child receives rewards for
appropriate and not maladaptive behavior.

**The AB Design**

Before examining valid small-
 designs, we look at a common but invalid way to evaluate the effectiveness of a therapy. Research concerning the effectiveness of a therapy should be incorporated into the treatment whenever possible. This seems like a fairly simple matter: Measure the frequency of the behavior that needs to be changed, then institute the therapy and see if the behavior changes. We can call this an **AB design**, where $A$ represents the baseline condition before therapy, and $B$ represents the condition after therapy (the independent variable) that is introduced. This design is used frequently in medical, educational, and other applied research, where a therapy or training procedure is instituted to determine its effects on the problem of interest. However, the **AB** design (Campbell & Stanley, 1966) fails completely as a valid experimental design and should be avoided. It fails because changes occurring during treatment in the $B$ phase may be caused by other factors that are confounded with the factor of interest. The treatment might produce the change in behavior, but so could other sources that the researcher is not aware of or has failed to control. We cannot conclusively establish that change resulted from the therapy because of a lack of control comparisons. A confounding variable might have produced the change in the absence of an independent variable. Remember, confounding occurs when other variables are inadvertently varied with the primary factor of interest—in the case of the little girl, the therapy. It is crucial to control carefully the potential confounding variables, so that the primary one is producing the effect. This is impossible in the **AB** design, since the therapist-researcher may not even be aware of the other variables.

A standard solution to the problem involves a large-
 design. We have two groups to which subjects have been randomly assigned. One, the experimental, receives the treatment; the other, the control, does not. If the experimental condition improves with the therapy and the control does not, we may conclude that the treatment and not some extraneous factor produced the result. In the case of therapy on an individual, such as the case of the tantrum-throwing girl, there usually is no potential control group and only one subject in the experimental “group.” Since a large-
 design depends on having a substantial number of subjects in the experimental and control conditions, it is inappropriate for use in evaluating many therapeutic situations.
The ABA or Reversal Design
As an alternative to the flawed AB design, the experimenter may reverse the conditions after the phase to yield an ABA design, which is also called a reversal design. The second A phase in the ABA design serves to rule out the possibility that some confounding factor influenced the behavior observed in the B phase. Returning the conditions of the experiment to their original baseline level, with the independent variable no longer applied, allows the experimenter to determine if behavior returns to baseline level during the second phase. If it does, then the researcher can conclude that the independent variable effected change during the B phase. This generalization would not apply if a confounding variable happened to be perfectly correlated with the independent variable. Such a situation is unlikely. Here we consider an example of a reversal design.

Hart, Allen, Buell, Harris, and Wolf (1964) investigated the excessive crying of a 4-year-old nursery school pupil, Bill, who otherwise seemed quite healthy and normal. The crying often came in response to mild frustrations that other children dealt with in more effective ways. Rather than attribute his crying to internal variables, such as fear, lack of confidence, or regression to behavior of an earlier age, the investigators looked to the social learning environment to see what reinforcement contingencies might produce such behavior. They decided, with reasoning similar to that already discussed in the case of our hypothetical little girl, that adult attention reinforced Bill’s crying behavior. The researchers set about testing this supposition with an ABA (actually ABAB) design.

First, they needed a good measure of the dependent variable, crying. The teacher carried a pocket counter and depressed the lever every time there was a crying episode. “A crying episode was defined as a cry (a) loud enough to be heard at least 50 feet away and (b) of 5 seconds or more duration.” At the end of each nursery school day, the teacher totaled the number of crying episodes. We could, perhaps, quibble some with this operational definition of a crying episode (did the teacher go 50 feet away each time to listen?), but let us assume it is valid and reliable.

In the initial baseline of phase A, Bill received normal attention by the teacher to his crying. During the 10 days of the first baseline period, the number of crying episodes was between 5 and 10 per day, as shown in the leftmost panel of Figure 9.7, where the frequency of crying episodes on the ordinate is plotted against days on the abscissa. For the next 10 days (the first B phase), the teacher attempted to extinguish the crying episodes by ignoring them, while rewarding Bill with attention every time he responded to minor calamities (such as falls or pushes) in a more appropriate way. As shown in Figure 9.7, the number of crying episodes dropped precipitously, so that there were between zero and two during the last 6 days of the first B phase. This completes the AB phase of the design; once again, we cannot be certain that the reinforcement contingencies were responsible for Bill’s improved behavior. Perhaps he was getting along better with his classmates, or his parents were treating him better at home. Either of these things (or others) could have improved his disposition.

To gain better evidence that the reinforcement contingencies changed Bill’s behavior, the investigators returned to the baseline: Bill was again reinforced for crying. At first, he was rewarded with attention for approximations to crying (whimpering and sulking); after crying had been established again, it was maintained with attention to each crying episode. As the third panel in Figure 9.7 shows, it took only 4 days to reestablish crying. This led to the conclusion that the reinforcement contingencies, and not
any number of other factors, were responsible for the termination of crying in the first $B$ phase. Finally, since this was a therapeutic situation, the investigators instituted a second $B$ phase similar to the first, in which Bill’s crying was once again extinguished.

In this investigation, no inferential statistics were employed to justify the conclusions drawn. Rather, with good control of the independent variable and repeated measures on the dependent variable, the differences between conditions in this experiment were striking enough to decrease the need for inferential statistics. Use of $ABA$ small-$n$ designs can allow powerful experimental inferences.

**Alternating-Treatments Design**

As in the standard within-subjects experiments, small-$n$ experiments often include carryover effects that prohibit use of the reversal design. If the treatment introduced in the $B$ phase has long-term effects on the dependent variable, then reversal is impractical. Furthermore, the experimenter may want to obtain several samples of the subjects’ behavior under the same independent variable or under several independent variables. There are a number of ways to solve these problems, but we consider just two.

Rose (1978) used what could be called an $ACABCBCB$ design, where $A$ phases refer to baseline conditions, and $B$ and $C$ phases represent different levels of the independent variable(s). When presentation of different levels of the independent variable alternate, we have an alternating-treatments design. Rose wanted to determine the effects of artificial food coloring on hyperactivity in children. Two hyperactive 8-year-old girls were subjects. They had been on a strict diet, the Kaiser-Permanente (K-P) diet (Feingold, 1975), which does not allow foods containing artificial flavors and colors and foods containing natural salicylates (many fruits and meats). On the basis of uncontrolled case studies ($AB$ designs), Feingold reported that the K-P diet reduced hyperactivity.

Rose’s $A$ phase counted the behavior of the two girls under the ordinary K-P diet. The $B$ phase examined another kind of baseline. It involved the introduction of an oatmeal cookie that contained no artificial coloring. The $C$ phase included the level of interest of the independent variable: oatmeal cookies containing an artificial yellow dye. Rose chose this artificial color because it is commonly used in the manufacture of

![Figure 9.7](image-url)
foods, and it had the additional benefit that it did not change the taste or appearance of the cookies. (When asked to sort the cookies on the basis of color, judges were unable to do so systematically with regard to the presence of the dye.) The subjects, their parents, and the observers were blind to when the children ate the dye-laced cookie. Various aspects of the two girls’ behavior were recorded during school by several different observers. One dependent variable that Rose measured was the percentage of time that the girls were out of their seats during school. Rose found that the girls were most active during the C phases, when they had ingested a cookie with artificial coloring in it. Rose also noticed that there was no placebo effect. That is, the percentage of time out of their seats was essentially the same during the A phases (no cookie) and the B phases, in which the girls ate cookies without artificial coloring. So, Rose concluded that artificial colors can lead to hyperactivity in some children.

**Multiple-Baseline Design**

Rose’s extension of the reversal design allows an experimenter to examine the effects of more than two levels of the independent variable. However, the extension does not permit experiments involving independent variables that are likely to have strong carryover effects. The multiple-baseline design, illustrated in Figure 9.8, is suitable for situations in which the behavior of interest may not reverse to baseline levels (i.e., when there are permanent carryover effects).

Two features of the multiple-baseline design are noteworthy. First, notice that different behaviors (or different subjects) have baseline periods of different lengths prior to the introduction of the independent variable. The baseline periods are to the left of the vertical lines, and the treatment periods, in which the independent variable has been introduced, are to the right of the vertical lines. Using such a design in the case of Bill (described earlier) might involve a continual baseline monitoring of some other unwanted behavior (say, picking fights) when the extinction period for crying was introduced. Then after several days, perhaps, the extinction procedure could be applied to the fighting behavior. Behavior A in Figure 9.8 could be crying, and

![Figure 9.8](image_url)

**An Outline of the Multiple-Baseline Design.** Different people (between subjects) or different behaviors (within subjects) have baseline periods of different lengths. The vertical lines indicate when the independent variable (the treatment) was introduced.
Behavior B could be fighting. Fighting behavior occurs under baseline conditions while the crying behavior is treated. If the untreated behavior holds steady prior to the introduction of the independent variable and then changes afterward, we can assume that the independent variable alters the behavior and not something else that also happens to change during the observation period. Suppose, however, that crying and fighting typically go together. So, the treatment of one behavior could influence the occurrence of the other. If Bill’s fighting decreased as crying extinguished, then we could not attribute the changes in one of the behaviors to the independent variable.

This problem leads us to the second important feature of the multiple-baseline design. The multiple-baseline design can be used as a small-n equivalent of the between-subjects design. As shown in Figure 9.8, instead of several behaviors being monitored as in a within-subjects design, different people can be monitored for different periods prior to the introduction of the independent variable. This type of multiple-baseline design, as is true of the ordinary between-subjects design, should be appropriate for situations in which the independent variable will have strong carryover effects. The between-subjects multiple-baseline procedure is also appropriate for cases where target behaviors are likely to be influenced by each other, such as could have occurred in our hypothetical experiment concerning Hart and coworkers’ subject, Bill.

An experiment by Schreibman, O’Neill, and Koegel (1983) nicely illustrates the between-subjects form of the multiple-baseline design. Schreibman and her co-workers wanted to teach behavior-modification procedures to the normal siblings of autistic children so that they could become effective teachers of their autistic siblings. Autism is a behavior disorder of unknown origin. It is characterized by impoverished social behavior, minimal language use, and self-stimulation of various kinds. For each of three pairs of siblings—one normal (mean age was 10 years) and one autistic (mean age was 7 years)—several target behaviors, such as counting, identification of letters, and learning about money, were chosen for the normal sibling to teach the autistic sibling. Since the normal siblings had to learn correct behavior-modification techniques, such as reinforcement for appropriate responding, the experimenters first recorded baseline measures of the normal siblings’ use of correct behavior-modification techniques and the correct performance of the target behaviors by the autistic children. The baseline data for each pair of children appear left of the vertical lines in Figure 9.9. Since learning behavior-modification techniques is likely to influence a wide variety of behaviors of the teacher and the pupil (the normal and autistic child, respectively), a multiple-baseline design across pairs of children was used. Changes in the behavior of normal and autistic children after the normal siblings were trained to use behavior-modification procedures appear right of the vertical lines in Figure 9.9. Correct performance by both children in each pair increased after the beginning of training. Schreibman and associates concluded that the training, and not some other confounding factor (such as changes resulting from being observed), altered the behavior.

Note the data points represented by plus signs and bull’s-eyes. These symbols show the children’s behavior in a setting that was entirely different from the training room, one in which the children did not know they were being monitored by the experimenters. Behavior in this generalization setting was very similar to the behavior in the training room, so the treatment program was effective in making general changes in the children’s behavior.
The operant-conditioning research designs described are representative of the powerful research techniques developed by Skinner and his followers. Careful control has provided an enormously valuable database for psychology. As we have seen, the procedures have been used in clinical settings with substantial success. The interested reader will see how broadly the techniques have been applied in Kazdin (2001). The small-\(n\) procedures of operant analysis are important tools for psychologists who wish to understand thought and behavior.

\[\text{Figure 9.9}\]

The normal siblings’ use of correct behavior-modification procedures, and the autistic children’s appropriate responses. The baseline period is to the left of the vertical line for each pair of children. The pluses and bull’s-eyes show behavior in a generalization setting (Schreibman et al., 1983, in *Journal of Applied Behavior Analysis*, 16, p. 135. Copyright 1983 by the Society for the Experimental Analysis of Behavior, Inc.) Reprinted with permission of Blackwell Publishing, Ltd.

The operant-conditioning research designs described are representative of the powerful research techniques developed by Skinner and his followers. Careful control has provided an enormously valuable database for psychology. As we have seen, the procedures have been used in clinical settings with substantial success. The interested reader will see how broadly the techniques have been applied in Kazdin (2001). The small-\(n\) procedures of operant analysis are important tools for psychologists who wish to understand thought and behavior.
**Changing-Criterion Design**

The **changing-criterion design** involves changing the behavior necessary to obtain reinforcement. For example, a rat may have to press a lever for food reinforcement five times for several minutes, and then the criterion behavior could change to seven lever presses to get the reinforcement. This procedure could then repeat with several other criteria. Here the independent variable is the criterion behavior necessary to obtain the outcome, and the underlying logic is similar to that of other small-n designs. If behavior changes systematically with the changing criteria, then we assume that the criteria are producing the change.

Therapists use a changing-criterion design in a variety of behavioral therapy situations. Kahng, Boscoe, and Byrne (2003) used such a design to increase food acceptance in Clara, a 4-year-old girl. She would drink out of a bottle, but she would not eat food. The therapist used a clever procedure of having escape from the meal contingent on a certain number of bites of food. Further, if Clara accepted a bite of food, she received praise and then later in therapy a Blues Clues token that she could play after the meal. The therapist increased the criterion number of bites to escape the meal setting for her favorite food (applesauce). She gradually accepted applesauce more readily with increases in the criterion. Acceptance of other foods gradually increased as well. At the end of therapy she met a 15-bite criterion of all foods in about 16 minutes. At a follow up meal 6 months after therapy, she ate more than 90 percent of 50 bites in 10 minutes.

McDougall and his associates (McDougall, Hawkins, Brady, & Jenkins, 2006; McDougall & Smith, 2006) have developed interesting variations of the changing-criterion design. The **range-bound changing criterion** involves criteria that specify upper and lower bounds of the target behavior. Thus, a child who needed more aerobic exercise might have criteria that specify the minimum and maximum amount of running for each exercise period. The **distributed-criterion design** shares features of the multiple-baseline design and the alternating-treatments design. In this design, a child who did not play well with others might have criteria for the amount of time spent in solitary and in social play during recess. These percentages and criteria are distributed across two or more behaviors and could be changed across time to increase the desired behavior (social play in this case). A lesson to take away from McDougall’s work is that the various designs are powerful ways of changing behavior and that combining these designs can enhance the therapeutic situation.

### FROM PROBLEM TO EXPERIMENT

#### THE NUTS AND BOLTS

**Problem**  The Partial Reinforcement Extinction Effect

To produce instrumental learning (operant conditioning), we follow with a reinforcement stimulus the behavior that we are interested in having the animal learn. The animal soon learns that the reward is forthcoming in the situation if the appropriate response is emitted. For example, suppose we want to teach an animal to learn a maze. The simplest sort of maze is the straight alley, which is composed of a start box where the animal is placed, an alley through which the animal runs when the start box door is opened, and the goal box where
the animal is reinforced. The reinforcement is typically food, and usually the animal has been deprived of food prior to the experiment. The dependent variable is running speed or time to run the straight-alley maze. Often the animal’s speed in each section of the runway is measured so that the experimenter finds speeds for its leaving the start box, traversing the alley, and approaching the goal. Learning is indicated by the fact that after a number of trials the rat’s speed increases (the latency decreases). At first, the rat dawdles along, but on later trials, it really hustles.

**Problem** How is learning affected by the amount of reinforcement?

Suppose we now wanted to ask a straightforward question about learning in this situation: How is learning affected by the amount of reinforcement? Intuitively, you might expect that learning would increase as the amount of reinforcement increases. But if you read the first part of this chapter carefully, you should realize that this depends on how “amount of reinforcement” and “learning” are defined. We could vary the amount of reinforcement by varying the percentage of trials on which subjects receive reward or by varying the magnitude of reward after each trial. We could also measure learning in several ways: one might be running speed, another might be judgments of persistence (by humans, of course, Svartdal, 2003), or we could measure resistance to extinction. The latter measure, discussed in the “Introducing the Variables” section, is found by seeing how long after training an animal will continue running a maze when it no longer receives reinforcement.

**Problem** What is the effect of percentage of reward on resistance to extinction?

Let us confine our interest to the case in which we vary the percentage of rewarded trials. Our experiment has now become more manageable.

We vary the percentage of trials on which the animals receive rewards for running the maze (the independent variable), and we measure the time it takes the animals to run the maze and their resistance to extinction (or running speeds during extinction training).

In many experiments such as these, researchers have found that resistance to extinction is generally greater the **smaller** the percentage of trials during which the animal receives reinforcement during training. If an animal receives **continuous reinforcement** (i.e., is reinforced after every trial), its running behavior will extinguish much more rapidly when reinforcement is withdrawn than will animals that receive reinforcement on only some percentage of the acquisition trials. In general, the smaller the percentage of reinforced trials, the greater will be the resistance to extinction (i.e., the faster the animal will run when reinforcement is withdrawn). The fact that infrequent reinforcement will lead to greater persistence in responding than continuous reinforcement is called the **partial reinforcement extinction effect (PREE)**. Several explanations of it have been proposed (Amsel, 1994; Capaldi, 1994).

A number of variables may contribute to the typical PREE. When rats are rewarded on only some proportion of trials, a number of factors may vary...
simultaneously. One factor is the number of nonrewarded trials (or N-trials) that precede a rewarded (or R) trial. Another factor is the number of transitions from nonrewarded to rewarded trials (or N–R transitions) during the course of partial-reinforcement training. A third factor is the number of different N-lengths (or number of different sequences of nonrewarded trials preceding a rewarded trial) during partial reinforcement. All these variables could be (and have been) examined; let us consider the first by way of an experiment in animal-learning research.

**Hypothesis**  
Resistance to extinction will increase with increases in the number of nonrewarded trials that precede a rewarded trial (the N-length).

Basically, we want to design an experiment in which the number of nonrewarded trials would be varied before a rewarded trial. There could be three N-lengths of one, two, and three nonrewarded trials before a rewarded trial, with the hypothesis being that resistance to extinction should increase with N-length. The greater the number of nonrewarded trials, the faster the rats should run during extinction.

A simple straight alley is used as the training apparatus, and the time for the rat to run the maze is the dependent variable. Do we want to use a within-subjects or between-subjects design? If we use a within-subjects design, we have to counterbalance the three schedules of reinforcement. But even if we do this, there is a serious problem of a carryover effect, or the effect that training rats under one schedule has on training them on the next.

When just one response is examined, such as running in a straight alley, a between-subjects design is used. However, it is possible to avoid a carryover problem by having the subjects learn different responses under different schedules of nonrewarded trials. Animals could receive one N-length in a straight alley painted black and a different N-length in a white alley. This kind of within-subjects experiment on the PREE has often yielded the surprising result of a reversed PREE, in which the animal shows greater resistance to extinction for the response that has the larger percentage of reinforced trials (for a discussion of some of the issues involved in a within-subjects PREE, see Rescorla, 1999). In a single experiment with human participants, Svartdal (2000) observed the usual PREE in a between-subjects comparison and a reversed PREE in a within-subjects comparison. If you are interested in tying together an understanding of design effects, contrast effects, and partial reinforcement, we suggest you examine the work by Rescorla (1999) and Svartdal (2000) and design your own experiment.

Stable results in a between-subjects partial-reinforcement situation could probably be achieved with only 15 subjects in each of three groups. Before the experiment is begun, it is usual to pretrain the animals to get them used to the experimental situation. This reduces the amount of within-subjects variability caused by extraneous factors, such as fear of being handled by the experimenter. Thus, for several days, the animals are handled for an hour or so each day by the experimenter. The rats should also be placed in the goal box with food pellets in the food dish to ensure that they will eat the pellets. Otherwise, as you might readily suppose, the pellets are unlikely to serve as a reinforcer. Finally, the rats should be placed in the straight alley and allowed
to explore it for a few minutes on each of several days before actual testing. This is to ensure that they will not be frightened when they are placed in it for testing. On each trial of the experiment proper, the experimenter takes the rat from its home cage and places it in the start box. The start box door is opened, which starts a timer, and the rat moves down the alley to the goal box. Near the goal box, the rat passes through a photoelectric beam, which stops the timer. When the rat enters the goal box, the goal box door is closed, so that the rat cannot return to the alley. Typically, the rat is confined to the goal box for a constant period of time in all conditions—say, 30 seconds. Then the rat is placed in a separate cage to await the next trial. In this experiment, the independent variable is the number of nonrewarded or $N$-trials (one, two, or three) preceding a rewarded or $R$-trial. This is straightforward, since it is easy either to provide or not to provide food when the rat runs the maze. The only tricky aspect is that the experimental procedure confounds the number of nonrewarded trials with the amount of rewarded trials that the rats receive during a series of tests. The rats with greater $N$-lengths receive less reward. One way to correct this confounding is to provide subjects in conditions with $N$-lengths of two and three with intertrial reinforcements. These are simply periods when rats are given rewards between trials in the neutral cage. The rewards are not dependent on the instrumental response.

The rats should be given a number of days of training, perhaps 10, to ensure that they learn their particular schedule of reinforcement. Twelve trials per day would be an appropriate number. After 10 days of learning, extinction training is introduced. This consists of simply running the rats at 12 trials per day for perhaps 4 days with no reward at all and measuring the time the rats take to run the maze. This phase of the experiment is critical, since we want to ascertain the effect of the training schedules on resistance to extinction. But a problem enters here. It is common during extinction for at least some rats to simply stop running. Either they refuse to leave the start box, or they stop halfway down the alley. What happens to our dependent measure in cases such as this? The convention adopted to avoid this problem is to allow the rats a fixed amount of time to traverse the alley and to remove them and begin the next trial if they fail to beat the cutoff. A limit of 90 seconds is often used; if a rat has not made it a few feet down the maze in 90 seconds, it is unlikely that it will make it at all. Thus, the experimenter simply removes the animal, records its time as 90 seconds, and places it in the neutral cage in preparation for the next trial. Since different schedules of training sometimes produce lopsided or skewed distributions of running times, it is often necessary to use the median time for each animal rather than the mean to eliminate the effect of a few extremely long times. (See Appendix B for a discussion of medians.)

The basic purpose of the experiment is to see whether the number of nonrewarded trials produces greater resistance to extinction. In other words, if subjects receive greater $N$-lengths, will they run faster when given no reward during extinction? In an experiment similar to the one described here, Capaldi (1964) found that greater $N$-lengths were associated with greater resistance to extinction.
SUMMARY

1. The study of animal learning and behavior has identified two basic types of conditioning. In classical (or Pavlovian or respondent) conditioning, a neutral stimulus, such as a light or tone, precedes an unconditioned stimulus that produces an automatic or unconditioned response.

   After a number of such pairings, the originally neutral stimulus produces the response if there is a contingent (predictive) relation between the neutral stimulus and the unconditioned stimulus.

2. In instrumental (operant) conditioning, a particular behavior that is followed by a reinforcing stimulus will increase in frequency. A positive reinforcing stimulus (the familiar rewards) increases the frequency of the response that produces it. A negative reinforcing stimulus increases the frequency of the response that removes it. A response that is followed by a punishing stimulus decreases in frequency.

3. In all experimental research, whether with humans or other organisms, a fundamental question is whether to use the same organisms in each condition of the experiment (a within-subjects design) or to use different organisms in the different conditions (a between-subjects design).

4. Within-subjects designs are preferred when they can be used, because they minimize the amount of variability caused by differences among subjects. Also, within-subjects designs employ fewer subjects than between-subjects designs, though, of course, it is necessary to test each individual for longer periods.

5. The primary danger in within-subjects designs is that of carryover effects, the relatively permanent effects that testing subjects in one condition might have on their later behavior in another condition. In such cases, it is necessary to use between-subjects designs, even though more subjects will be needed and subject variability is less controllable.

6. The choice of an experimental design may in some instances strongly affect the outcome of an experiment. For example, stimulus intensity appears to play little role in classical conditioning when manipulated between subjects but a great role when manipulated within subjects. Thus, the type of experimental design chosen can sometimes be critical.

7. In within-subjects designs, it is necessary to counterbalance conditions or to vary the conditions in a systematic way so that they are not confounded with time of testing. If conditions are not counterbalanced, then time-related effects, such as fatigue or practice, rather than manipulation of the independent variable may account for the results. It is also necessary to counterbalance in between-subjects designs across variables that are not of central interest. One quite useful counterbalancing scheme is the balanced Latin square design, in which each condition precedes and follows every other one equally often.

8. The traditional large-\(n\) research methodology is often inappropriate in applied settings, where there is only one subject. Often an \(AB\) design is used as a small-\(n\) design. A baseline of behavior is established (the \(A\) phase), and then some treatment is imposed (the \(B\) phase).

   The conclusion that changes in behavior during the \(B\) phase resulted from the treatment is faulty because other variables may be confounded with the treatment (such as practice or fatigue effects).

9. The \(ABA\) design is a powerful alternative to the \(AB\) design. The second \(A\) phase, introduced after the \(B\) (treatment) phase, removes the treatment to determine whether any changes observed during the \(B\) phase were caused by the independent variable or by confounding factors. The alternating-treatments design permits an examination of several independent variables or of independent variables with more than two levels.

10. The multiple-baseline design is another small-\(n\) design in which different behaviors or different people receive baseline periods of varying lengths prior to the introduction of the independent variable. For use within subjects, this design is preferable to the \(ABA\) design if the independent variable has strong carryover effects.
Key Terms

AB design  
ABA (reversal) design  
ABAB design  
ABBA design  
alternating-treatments design  
asymmetrical transfer  
balanced Latin square design  
between-subjects design  
blocking  
carryover effect  
changing-criterion design  
classical conditioning  
conditioned response (CR)  
conditioned stimulus (CS)  
contingency  
continuous reinforcement  
counterbalancing  
discriminative stimulus (SD)  
distributed-criterion design  
extperimental extinction  
fatigue effect  
instrumental conditioning  
large-n designs  
matched-groups design  
multiple-baseline design  
negative contrast effect  
negative reinforcing stimulus  
null contingency  
operant conditioning  
partial reinforcement extinction effect (PREE)  
positive contrast effect  
positive reinforcing stimulus  
practice effect  
psychoneuroimmunology  
pseudoconditioning  
punishment  
random-groups design  
range-bound changing criterion  
respondent conditioning  
reversal (ABA) design  
shaping  
simultaneous contrast  
small-n designs  
split-litter technique  
unconditioned response (UR)  
unconditioned stimulus (US)  
within-subjects design

Discussion Questions

1. Discuss the advantages of within-subjects designs. What complications and problems are entailed by using a within-subjects design?
2. Discuss the advantages and disadvantages of using a between-subjects design.
3. In each of the following cases, tell whether it would be best to examine the independent variable in a within-subjects or a between-subjects design. Justify your answer in each case.
   a. A social psychological study of helping, in which the researchers are interested in how group size affects whether or not an individual will help someone else in the group.
   b. A study of the effect of varying loudness of a tone in measuring how quickly people can respond to the tone.
   c. An experiment designed to answer the question of whether the color of a woman’s hair affects the likelihood that she will be asked out for dates.
   d. A study in which three different training techniques are compared as to their effectiveness in teaching animals tricks.
4. Tell what a balanced Latin square is and explain why it is a preferred counterbalancing scheme in many situations. Draw two balanced Latin squares similar to those in Tables 9.1 and 9.2 for cases in which there are (a) three conditions and (b) four conditions.
5. The results of some experiments described in this chapter showed different effects of an independent variable when it was manipulated between
and within subjects. Make a list of three variables for which you think between- and within-subjects designs would show the same effects, and provide two further instances in which you think the two types of designs would produce different results. Justify your reasoning in each case.

6. Discuss the sorts of confounding that may arise from the use of an AB design.

WEB CONNECTIONS

Explore the step-by-step presentation of “Between vs. within Designs” at:
http://academic.cengage.com/psychology/workshops/student_resources/workshops/between1.html

A helpful overview of learning is available at:
http://www.funderstanding.com/theories.cfm

A good discussion of research design can be found at:
http://trochim.human.cornell.edu_kb/expfact.htm

PSYCHOLOGY IN ACTION

Knowledge of Results as Reinforcement

People receive many reinforcements in the form of knowledge of results rather than as biological rewards, such as food pellets given to hungry rats in Skinner boxes. “That’s good” or “You’ve almost got it correct” are frequently given as feedback. So, in addition to being rewarded for approximations to a correct response, we also are told how close we are to a target response.

The following is based on a famous experiment done by Thorndike (1932). We provide a variant of his procedure that was suggested by Snellgrove (1981). Thorndike blindfolded subjects and asked them to draw lines that were 3 inches long. Little or no improvements in accuracy occurred when the subjects did not receive knowledge of results, but Thorndike found rapid shaping of behavior when knowledge of results was given: People were told “right” when they were within one-eighth inch and “wrong” when they were off by more than one-eighth inch. You will vary the type of feedback: Some participants will be told nothing, others will be told “good” when they are within one-eighth inch of 3 inches, and a third group will be told exactly how long a line they drew (to a sixteenth of an inch). You will need paper, pencils, a ruler, and a blindfold. Each person will receive 10 trials, with a single kind of knowledge of results. Record the accuracy (to a sixteenth of an inch) on each trial for each person, so you can compare the rate of progress across the 10 trials for each form of feedback.

You could use a within-subjects design, in which a person tries drawing different lengths of line (e.g., 2 inches, 3 inches, and 5 inches) under one of each of the feedback conditions. With a within-subjects design, you will need to counterbalance the type of feedback across line length, as well as the ordering of conditions throughout the experiment. This experiment is probably best done with a between-subjects design, because of relatively permanent carryover effects resulting from the different kinds of feedback.
If you use a between-subjects design, you should combine your efforts with classmates, so that you can have a large sample of participants in each feedback group.

Be sure to agree on a protocol, because all experimenters must treat the participants identically (except for the levels of independent variable) during testing. The timing for giving feedback must be the same for each feedback condition. Delayed knowledge of results usually yields better learning than immediate knowledge (e.g., Swinnen, Schmidt, Nicholson, & Shapiro, 1990), so do not confound how quickly feedback is supplied with the nature of the feedback. On each trial, measure and record performance, then give the appropriate knowledge of results before going on to the next trial.