CONTENTS SINGLE-CASE EXPERIMENTAL DESIGN

9 SINGLE- CASE EXPERIMENTAL DESIGNS

> *When you have the responsibility of making absolutely sure a given organism will engage in a given sort of behavior at a given time, you quickly grow impatient with theories c learning. Principles, hypotheses, theorems, satisfactory proof at the .05 level of significance . . . nothing could be more irrelevant. No one goes to the circus to see the average dog jump through a hoop significantly oftener than untrained dogs. B. F. Skinner*

In the early 1990s, a young girl who was enrolled in a special education class in a New England school stunned her teachers by revealing that she was being sexually abused by every member of her family (Palfreman, 1993). Serious doubts arose concerning the truth of her story, however, because the girl did not report the abuse directly. She was autistic and could not speak normally. Instead, she had typed her message with the help of another person using a new technique called facilitated communication. Although her family denied the allegations, the authorities initiated a court case to see whether the child should be removed from her home.

Facilitated communication had just recently been introduced as a breakthrough in treating autism. Autistic children exhibit severe social withdrawal, gravely impaired communication, and restricted and ritualized behavior. Most autistic children score in the subnormal range on measures of intellectual development. However, the advocates of this new form of communication claimed to have discovered that autism is a disorder that traps normal, even gifted minds inside poorly functioning bodies, which are physically unable to produce speech or make signs. With the help of facilitated communication, they believed, many autistic children would be shown to have normal intellectual abilities.

Facilitated communication is done with a trained adult, called the facilitator, and a computer. With the help of the facilitator, the child types out messages with one finger on the keyboard. The facilitator steadies the child's hand and helps keep the child focused by placing

his or her hand on the child's arm. This gesture also is thought to provide the child with much-needed emotional support.

The new technique produced astounding results. Autistic children who could hardly communicate before began expressing their thoughts and feelings with amazing clarity, using advanced vocabulary, and with good grammar and punctuation. Autistic children took advanced mathematics and English courses with their facilitators. Some typed out messages of relief at having been set free from the prison of their autism. The method, which was hailed as a miracle that would enable autistic children to lead productive lives, was introduced into many school systems.

But not everyone accepted these findings. The method was not based on solid research and there was an unresolved question about the nature and extent of the facilitator's influence over what was being communicated by the child. In fact, this issue became the focus of the court case. Was the claim that the child was being sexually abused coming from her or from her facilitator? Could the facilitator be controlling what the child typed by guiding her hand rather than just steadying it?

The court agreed to a scientific experiment to answer this question. The experiment was designed and carried out by Howard Shane, an expert on the communication of handicapped people. Because the court needed to resolve the question of whether the facilitated communications revealed the thoughts of the child or her facilitator, by necessity, the experiment had to involve only one subject, the autistic child. In addition, it was vital that the experiment be internally valid. Any error in its conclusion would lead to tragedy, either needlessly separating a child from loving parents, or reuniting her with abusive ones.

Shane used an ingenious experimental method to test the validity of facilitated communication. The experiment consisted of a series of trials on each of which the child was shown a picture of an everyday object, for example, a key or shoes, and asked to type the name of the object with the aid of her facilitator. The facilitator was shown the same picture as the child on half the trials; on the other trials, the facilitator saw a different picture than the child. Neither the child nor

the facilitator could see the picture shown to the other person. The two types of trials were presented in random order and a record was kept of what was typed.

The logic of the experiment is clear. If the facilitator is merely helping the child to express her own thoughts, the words typed should not be influenced by what the facilitator sees. However, if the facilitator is controlling the content of the typing, then, if they see different pictures, the child should type the name of the object that the facilitator sees.

The results were definitive. On every trial in which the child and facilitator saw different pictures, the child named the object seen by the facilitator. ! Statistical tests were not needed. The probability that this pattern of typing occurred due to uncontrolled variables was negligible. The conclusion was inescapable. The child was not communicating her own thoughts. Accordingly, the charges were dropped and the family reunited.

Other researchers subsequently did their own tests of facilitated communication with similar results to Shane's. As a result, many experts no longer consider facilitated communication a breakthrough for understanding autism (Delmolino & Romanczyk, 1995, and Jacobson, Mulick, & Schwartz, 1995) and some programs using facilitated communication have been discontinued. Children previously placed in advanced classes with facilitators have been removed from them and put back into classes appropriate to their abilities. Shane's simple experiment, which tested only one subject, saved a young girl and her family. It also led to a clearer understanding of a once promising method for reaching autistic children.

Shane's experiment is a dramatic example of the need in psychology for *single-case* or $n = 1$ *designs*. These designs were developed so that psychologists could study the conditions affecting behavior experimentally with single subjects. The research designs we have focused on previously in this book (e.g., Galton's correlational studies and Fisher's factorial designs) require many subjects, to permit random assignment of subjects to conditions and/or to compute the necessary correlations. Of course, these methods cannot be applied to single cases. So researchers using single-case designs have had to

devise different techniques for dealing with uncontrolled variables. In this chapter, we discuss the logic and variety of these experimental designs.

9.1 OXO Single-Case Design

Previous chapters of this book have pointed to two advances in methods that have had a dramatic impact on how psychologists conduct research. They are Fisher's invention of randomized experimental designs (see Chapters 6 and 8) and the development of statistical controls (see Chapters 5 and 8). So popular are these methods that, if you were to pick up a major psychology research journal and browse through the articles in it, you would discover that randomized groups, statistical controls, or both, would be used in the great majority of them.

If you were to look through a major journal in the physical sciences, however, you would not find these methods to be anywhere near as popular as they are in psychology. In fact, undergraduate curricula in physics and chemistry usually don't even consider these methods. The reason is that courses in the statistical aspects of experimental design are unnecessary in these fields. In physics and chemistry, the basic experiments simply follow Mill's experimental method exactly; everything is controlled except for one condition that is varied.

Consider a simple experiment in chemistry. Suppose a chemist wants to test the hypothesis that if a drop of Chemical A is added to a beaker of Chemical B there will be a reaction that releases heat. The experiment would be straightforward: first measure the temperature of the chemicals to make sure that both are at room temperature, then add the drop of Chemical A to B, and finally measure the temperature of the mixture again. A sudden increase in temperature would confirm the hypothesis.

The results of this experiment would be convincing because: the time of the mixing is determined by the experimenter; external conditions are well controlled; the treatment takes only a moment to complete; the expected result occurs immediately after the chemicals are mixed; and the result is dramatic— there is an unmistakable rise in

temperature. The procedure could be replicated over and over with the same result.

This type of study can be diagrammed as:

OXO

where O stands for an observation and X stands for a treatment (Campbell & Stanley, 1963). The "subjects" in this experiment are objects or chemicals. The experiment involves repeated measures; the same "subject," in this case Chemical B, is measured twice, once before the treatment (a pretest) and again after the treatment, adding Chemical A (a posttest). This would be called a single-case design in psychology, since one "subject" rather than a group of subjects is observed.

The pioneers of psychological research patterned their experiments after this popular design in the physical sciences. Franklin's evaluation of Mesmer's magnetic cure (see Chapter 1) used the O X O single-case design. A patient was observed, given a fake magnetic treatment, and tested once again. Hermann Ebbinghaus (1895/1913) did his experiments on memory using a repeated measures design with himself as the only subject. He studied how variations in lists of nonsense syllables that he memorized affected his subsequent recall of them. Ivan Pavlov (1928), in his classical conditioning experiments, "first observed that a dog did not salivate at the sound of a bell, then he exposed the dog to many pairings of the bell with food, and finally he retested the dog to see whether it now salivated to the sound. John Watson, the behaviorist (1928/1972), used this design to demonstrate how a phobia could be created in a young child.

The O X O single-case design lost favor in psychology once it became apparent that it was inadequate for studying many problems of interest to psychologists. This design has problems, for example, when the treatment, X, is not one that produces immediate and dramatic effects, a common occurrence in psychological research. In evaluating an educational program, like Head Start, for example, the treatment, early education for children, would take months or years to administer, the effects would not be apparent immediately, and the effects would not be dramatic for every child—some would benefit,

others would not. Also, when a long time intervenes between the Os in the design, the threats of history (external events that affect the posttest measure) and maturation (changes in the subject over the course of the study) are problematic.

Psychologists have adopted the control group design with random assignment of subjects to conditions to control for these threats. This two-group design can be diagrammed as follows:

Experimental group: R O XO

Control group: R O O

where R indicates that subjects are randomly assigned to the groups, X, the treatment, is given only to the experimental group, and the control group receives no treatment. As we have discussed in previous chapters, this randomized design is considered by many psychologists to be the ideal method for psychological research.

However, some experimental psychologists do not use this control group design at all in their research. Instead, they argue that for the research problems of interest to them the single-case O X O design and similar designs are preferable to the randomized control group design. The views of these psychologists are a legacy of the pioneering advances in single-case methodology that B. F. Skinner made starting in the 1930s.

Skinner used the O X O single-case design, with important modifications, in his experiments on animal learning. His treatments involved giving a subject, usually a pigeon or rat, a reinforcer, food, for performing a particular type of response. The reinforcement produced immediate and dramatic effects, comparable to those found in experiments in physics and chemistry.

Skinner's success with the method inspired others to adopt his operant conditioning procedures and his experimental methods. During the 1950s and 1960s, researchers showed the effectiveness of behavioral techniques for treating phobias, autism, obsessive compulsive disorders, enuresis, addictions, anorexia and obesity, using a modified single-case O X O design. In 1960, Murray Sidman published *Tactics of Scientific Research,* which became the bible of single-case research

design. Nothing succeeds like success, and the single-case design was reestablished as a viable method in psychology.

Psychologists who advocate the use of single-case experimental designs in psychology do not see them as a replacement for the randomized groups design. Instead, they consider the single-case experiment to be a superior method for certain applications and the only possible one for others. Before discussing modern developments in single-case designs, we will look at some of the problems with randomized control group designs that have led to the renewal of interest among psychologists in single-case designs.

9.2 Limitations Of Randomized Control Group Designs

9.2.1 Getting Enough Subjects

Randomized control group designs require many subjects in each group. We learned in Chapter 6, for example, that an experiment must have about 60 subjects in each of two groups for the recommended power of 80% to detect a moderate-sized effect. It is sometimes difficult to gather such large samples. For example, to collect enough subjects for a study of depression with three treatment groups and one control group, researchers had to recruit participants from several different hospitals and clinics in three separate cities (Elkin et al., 1989). And depression is often called the "common cold" of psychological disorders. If the research focuses on rare conditions, like dissociative disorders or pica (compulsive eating of nonnutritive substances), finding enough subjects for a randomized design may be impossible.

Single-case research designs are naturally suited to the study of rare phenomena. In addition, they have proven popular among clinical psychologists who are interested in assessing the effectiveness of treatments for particular clients rather than in conducting large-scale studies involving many people.

9.2.2 Misleading Summary Statistics

Between-subjects designs typically use the mean and standard deviation as summary statistics to describe the experimental results. Individual scores of subjects usually are not reported. In fact, the *Publication Manual of the American*

TABLE 1 CHANGE SCORES ON THE HAMILTON DEPRESSION SCALE FOR INDIVIDUAL SUBJECTS DIAGNOSED WITH SAD: PRE- MINUS POSTTHERAPY DEPRESSION SCORES[1] (FROM ROSENTHAL ET AL., 1985)

Bright Light	Dim Light
S1 -1	S8 -5
S2 7	S9 -3
S3 8	S10 3.5
S4 13	S11 4.5
S5 14	S12 5
S6 14.5	S13 6
S7	S14 9
	S15 9.5
	S16 25
Mean = 10.6	Mean = 6.1

Patients classified as atypical depression were not included in this table.

Psychological Association recommends against reporting individual scores in research using group designs (1994, p. 15). Unfortunately, summary statistics may not accurately reflect what happens to some subjects in a study, giving a misleading picture of the effects of the experimental treatments.

To illustrate, let's look at an experiment on the effects of light therapy for patients suffering from seasonal affective disorder (SAD), a winter depression thought to be caused by reduced light during the winter months. Rosenthal et al. (1985) tested the effectiveness of daily exposure to bright, full-spectrum artificial light in reducing the symptoms of patients with this disorder. The subjects, all diagnosed with SAD, were randomly assigned either to bright light therapy or to a placebo treatment, dim light. Ratings were made on the severity of their symptoms before and after a week of treatment.

Table 1 shows the change scores for individual subjects on the measure of depression from pretherapy to posttherapy. Positive numbers indicate an improvement in symptoms over the one-week experimental period. The mean change scores show that, overall, bright light resulted in a greater reduction in depression (M - 10.6)

than dim light (M = 6.1). However, inspection of the results for individual subjects reveals considerable variation in the effectiveness of the treatment.

The change scores of subjects 2 through 6, who received the bright light, were close to the mean of 10.6 (7 to 14.5). For these subjects, the mean accurately reflects the effect of the treatment. But the results for S1 and S7 were exceptions; S1 got slightly worse during the therapy and S7 improved a remarkable 18.5 points. The mean change of 6.1 for the placebo group accurately reflects only the effect of the treatment for S10 through S15, not for the other three subjects; S8 and S9 got worse under the dim light, but S16 improved by 25 points, the largest improvement in the entire study. S14 and S15, who also received dim light, improved to an extent close to the mean improvement of the subjects who got the bright light. If only the summary statistics had been presented, these exceptions would not have been noted and an incomplete understanding of the effect of the therapy would have resulted.

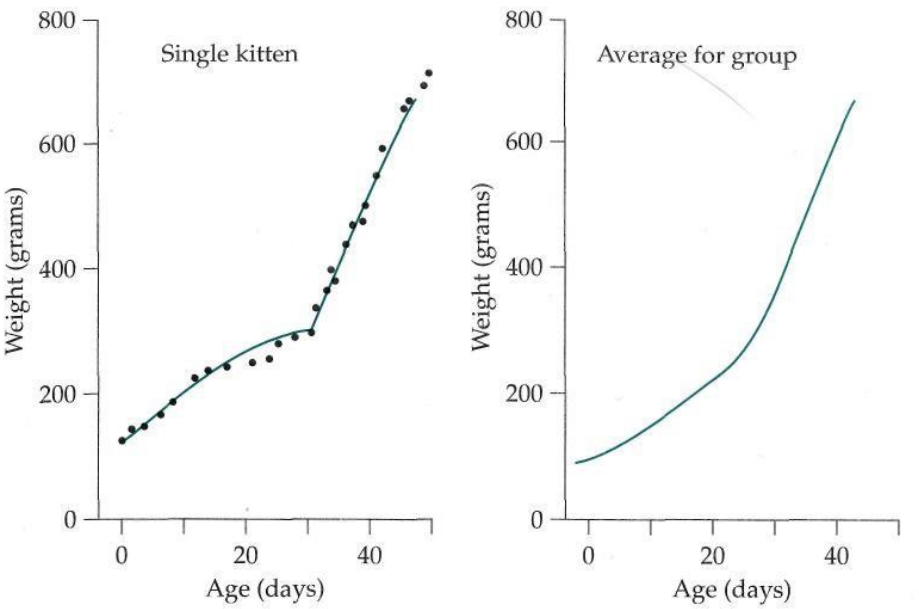


Figure 1. Growth curves for a single kitten and the average for a group of kittens. (From Bateson & Young, 1981)

In addition to burying the results for exceptional subjects, summary statistics can mask patterns that occur in all subjects. The data from

Bateson and Young's research on the growth of cats (1981, cited by Martin & Bateson, 1993) is a good example. Figure 1 plots weight as a function of age for one kitten in their study and for the combined data of several kittens. As you can see, the kitten whose weight is plotted in the individual record shows abrupt growth at about 30 days. This same rapid weight gain occurs in all kittens, but at different times. Because of the different timing of this increase, the growth curve for a group of kittens fails to reflect the abrupt change in weight that takes place in all kittens. Combining the data leads to the misleading conclusion that growth in kittens is smooth and uniform.

Single-case research avoids the problems of summary statistics by not using them. The results for each subject are reported separately, so it is always clear how the results apply to individual subjects.

9.2.3 Relevance

The randomized control group design is not suitable for certain research problems, such as testing the validity of facilitated communication for a particular child. Single-case designs are appropriate when the research focuses on the study of unusual problems (e.g., a rare disease), or unusual skills (e.g., photographic memory), or in certain invasive medical procedures (such as mapping the brain by stimulating various locations prior to surgery). Singlecase designs also are widely applied in diagnostic work when the researcher is interested in determining the cause of a particular patient's disorder. An allergist, for example, might conduct a series of O X O experiments to determine the L substances to which a patient is allergic.

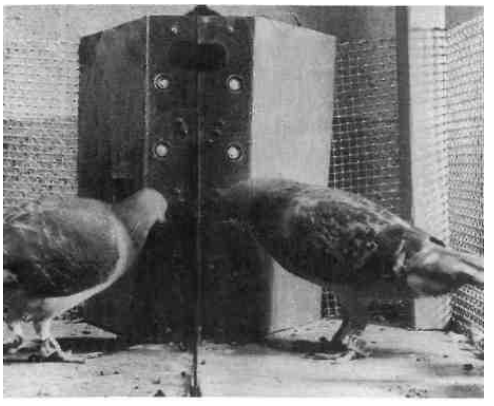
9.3 SKINNER'S BASIC EXPERIMENTAL DESIGN

The necessity of single-case research was impressed upon B. F. Skinner early in his career. In the 1940s, Skinner was involved in a project that required him, like the circus trainer in the quote at the start of this chapter, to be "responsible for making absolutely sure that a given organism will engage in a given sort of behavior at a given time." The project was not for a circus act, though. Skinner had obtained grants from General Mills and the U.S. Department of Defense to finance the training of pigeons for military combat in World War II! The quote at the beginning of this chapter was inspired by this project. Skinner's

plan was to have individual pigeons placed in the nose cone of missiles to guide them by pecking at a display showing the target. Whether the animals were accurate in performing this task was a life-or-death matter.

The standard control group research design was as useless to Skinner as it would be to the circus animal trainer. Skinner wasn't interested in demonstrating that pigeons trained according to a particular theory of learning would peck more accurately than untrained pigeons at an alpha level of .05. He had to find a method that would guarantee that each pigeon would perform accurately and consistently—and not under controlled conditions in the lab, but through all sorts of distractions, including heavy antiaircraft fire.

Skinner developed the basis for such a method while he was a graduate student in psychology at Harvard University. Skinner had entered graduate school without having taken a single psychology course. He had been an English major as an undergraduate, and apparently a good one. After reading some of his short stories, the famous poet Robert Frost suggested that he try writing as a career. Skinner spent a disappointing year after graduation trying before giving up this plan and entering Harvard. Except for some study of physiology, Skinner's background in psychology before graduate school was limited to reading Pavlov's work on conditioned reflexes and Watson's writings on behaviorism. Because he was untrained in the standard research methods of the field, Skinner developed his own, fashioned after the single-case methods of these pioneers.



Pigeons in Skinner's lab learning to cooperate. The pigeons are reinforced if they simultaneously peck matching buttons.

In his 1956 essay entitled "A Case History in Scientific Method," Skinner described the steps and lucky breaks that led to his invention of the method of operant conditioning. As a graduate student, Skinner wanted to develop a method for studying the effects of environmental manipulations empirically, by trial and error, without having to advance fancy principles or hypotheses. He decided to first establish the behavior of interest, then expose the subject to a particular treatment, then look again at the behavior. Instead of randomly assigning subjects to groups to deal with uncontrolled variables, Skinner adopted Pavlov's strategy of controlling for sources of variability before measurement by carefully controlling the animal's environment. (Recall Pavlov's "tower of silence" from Chapter 3.) With such control, Skinner reasoned, the effects of changing aspects of the environment should be immediately apparent. Statistical analyses of the data would be unnecessary.

Skinner built an ingenious apparatus for his experiments and developed a unique method for recording the animal's behavior. As you are probably aware, this apparatus is the "Skinner box," a special cage equipped with a bar or lever that the animal presses to get a pellet of food. The experimenter can program the box's food magazine according to a preset "schedule of reinforcement." The animal might be given a pellet for each bar press, for example, or for every 50 presses, or once every 5 minutes.

Skinner used a kymograph, an instrument that records the rate of response as a line on a strip of paper, to visually display the rate of bar pressing of individual subjects in the Skinner box. In kymograph recordings, called *cumulative records,* each response of the animal moves the stylus up, so that the slope of the line is proportional to the rate of the animal's responding—the steeper the slope, the greater the frequency of bar pressing.

The format for Skinner's experiments, which tested the effects of different treatments (e.g., schedules of reinforcement) on bar pressing, is as follows:

Prior to the experiment, a food-deprived animal is conditioned to associate the click of the food magazine with food. Then the animal is placed in the apparatus and the experiment begins. A *baseline phase,*

during which bar pressing is not reinforced, comes first. Its purpose is to determine the animal's response rate prior to the treatment. Next, during the *treatment phase*, which follows the baseline, bar pressing is reinforced, either continuously or intermittently. In the third phase, *extinction*, reinforcement is discontinued, baseline conditions are reestablished, and bar pressing no longer operates the food magazine.

B. F Skinner with a rat in a Skinner box. The kymograph is shown at the bottom right.

Finally, in the last phase, the reinforcement is reintroduced. This study can be diagrammed as:

O	**XO**	**XO**	**XO**
Baseline	Treatment	Extinction 2nd Baseline	Treatment

where O stands for continuous observation of the subject's behavior for a period of time; X is the start of the treatment; and X (read "not X") represents the start of extinction, when the treatment is withdrawn and baseline conditions are reinstated. An alternative system of notation that is used in the literature on single-case experiments labels the baseline phases "A" and the experimental

phases "B" (see Barlow & Hersen, 1984). The design is known as an *ABAB design,* or a *reversal* or *withdrawal design.*

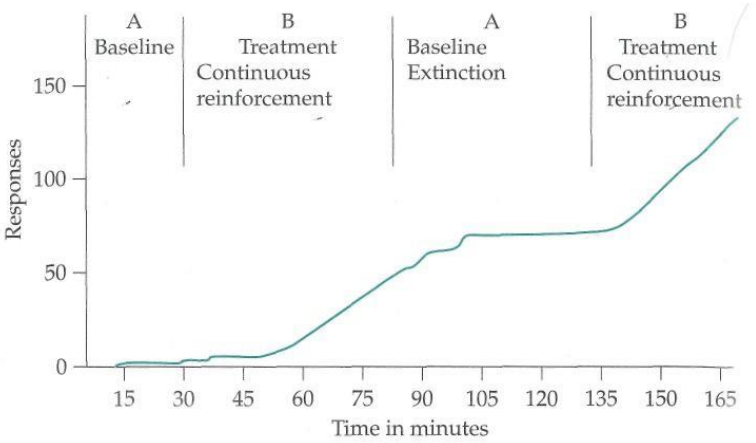


FIGURE 2 Cumulative record for an ABAB design

Figure 2 shows a cumulative record for an ABAB design. During the initial baseline the food-deprived subject presses the bar only a few times, as reflected in the flat cumulative record. In the first treatment phase, the slope of the cumulative record accelerates rapidly, indicating a steady high rate of bar pressing for this subject. The slope of the graph decelerates during extinction as the rate of bar pressing decreases. Finally, when the treatment is reintroduced, a high rate of bar pressing is quickly reestablished.

The results for this ABAB study demonstrate that the behavior of this subject is under the control of the experimenter. The rate of bar pressing is changed by giving or withholding reinforcement. Because everything other than the reinforcement is held constant throughout the experiment, there are no competing hypotheses to explain the pattern of the animal's behavior. Summary statistics and statistical tests are not needed, since the pattern of results is clear from visually inspecting the cumulative record.

Skinner's single-case design modified the O X O design used in chemistry in two ways. First, instead of a single pretest and posttest observation, Skinner used *continuous observations* of the subject before and during the treatment. Continuous observations are not needed in chemistry because variations in external conditions, which

are the only concern of the experimenter, can be controlled very well. But when the subjects are living creatures, changes in behavior occur spontaneously, without any apparent external causes. Continuous observations in the baseline phase enable the researcher to record the variability of the subject's behavior that is due just to uncontrolled external and internal events.

In addition, Skinner built replication into his design; once applied, the treatment is removed, then reintroduced later. In Skinner's research, the effect of the treatment extinguishes when reinforcement is discontinued, so it is possible for the experimenter to replicate the original experiment.

Skinner's modifications greatly improved the internal validity of the O X O design for research in psychology. The continuous observations in the baseline phase provide a check for the threats of history and maturation, since the effects of uncontrolled external events or changes in the subject would show up in the baseline record. The replication guards against any external event coincident with the treatment being confounded with the treatment. It is implausible that such coincidental events would occur repeatedly in the same experiment at the precise times that the experimenter introduces the treatments.

In addition to replicating the experiment for each subject, Skinner also replicated between subjects. To show that the treatment could be generalized to other subjects, as a rule Skinner would replicate the ABAB study for three more subjects. If the results were consistent for all four subjects, Skinner believed there was sufficient evidence to establish the generality of the effects of the treatment for the type of animal studied.

9.4 Replication In Modern Single-Case Designs

Skinner's ABAB design illustrates the distinctive features of many modern single-case experiments:

- The behavior of one subject is continuously observed for a period of time across different experimental conditions that are introduced by the experimenter.

- The subject's behavior during the treatment phase is contrasted with the same behavior during a baseline or no-treatment phase.
- The experiment is replicated within a single subject, for example, the AB sequence is repeated in the ABAB design.
- A graph showing all the observations made on each subject during each phase of the study is presented.
- The results can be analyzed with or without a statistical analysis.
- The experiment is replicated for additional subjects.

Earlier we noted that the internal validity of the single-case design is strengthened by replicating the treatment for a given subject. As the following discussion of modern variants of the single-case design illustrates, it is possible to replicate either sequentially or simultaneously:

1. In *sequential replication designs,* different experimental conditions are presented one at a time in sequence. The ABAB design is an example.

2. In *simultaneous replication designs,* several variations of the experiment are conducted at the same time on a single subject.

9.4.1 Sequential Replication Designs

ABAB Designs. Several variations of the ABAB design have been developed since Skinner did his original experiments. Modern experimenters sometimes elect to use partial replication rather than the complete replication that was Skinner's trademark. This can be done by dropping the first baseline, to give the *BAB design,* or by not reintroducing the treatment in the *ABA design.* Or two or more experimental conditions can be studied in sequence. For example, the *ABABCBC* design might be used to assess the effect of an experimental drug: *A* would be the baseline (no drug); *B,* the placebo drug (included to control for suggestion); and C, the active drug. This design actually combines two experiments in sequence: the ABAB design, to test the placebo effect, and the BCBC design, to test the active drug against the placebo.

Other designs of this general type can be created by using different numbers of treatments in varying orders (Barlow & Hersen, 1984).

In one variation of the sequential replication design, a single treatment is applied in varying degrees during the course of the experiment. A popular application of this *changing criterion design* is in evaluating the treatment for addictions, like smoking or drinking caffeine. In this application, the experimenter establishes a contract with the patient governing, say, the number of cups of coffee the patient can consume on any given day. The goal, or behavioral criterion, changes during the study. If the patient normally drinks 15 cups/day, the criterion for the first period might be 12 cups/day, then 10 cups/day for the second period, and so on, until the final goal is reached.
The experiment would be diagrammed as

$$O\ X_{12}\ O\ X_{10}\ O\ X_{8}\ O\ X_{6}\ O\ X_{4}\ O\ X_{2}\ O\ X_{0}$$

where X_8 indicates a criterion of 8 cups of coffee per day.

During the study, if the patient meets the criterion, by drinking the specified number of cups of coffee or less, a reward is given; otherwise there is a penalty. The treatment's effectiveness is demonstrated if the patient's behavior changes to match the criterion.

Alternating Treatment Designs (ATDs). The *alternating treatments design (ATD),* also called the *multi-element design,* differs by degree from the ABAB design. In both designs, the subject is exposed to a series of conditions, one after the other. Skinner's ABAB design had two experimental conditions, each presented twice in a fixed order, with many observations in each condition. In the ATD, more changes are made in the conditions and fewer observations are taken in each phase of the experiment. Shane's experiment on facilitated communication, in which he showed the child and the facilitator the same or different pictures, is an example of an ATD. In place of the fixed order of the ABBA design, the conditions in the ATD may be presented in a randomized order. For example, an ATD might have three conditions, each presented 30 times, in random order, with one observation in each condition. Using the O X system of notation, an ATD alternating randomly between three treatments (X,Y, and Z) could be diagrammed as:

OYOXOZ	OZOXOY	OZOYOX, . . . ,	OYOZOX
Block 1	Block 2	Block 3	Block n

The randomization here is within blocks; all three treatments are presented in a random order, then all three are presented again in a random order, and so on. Such *block randomization* guarantees that no one treatment is presented too many times in a row or left out for an extended period of time.

The ATD has two major advantages over the ABAB design. First, it usually involves many more replications of the basic AB experiment, providing an opportunity to demonstrate over and over that the subject's behavior can be controlled by the experimenter's choice of experimental conditions. Second, the order of the treatments can be randomized. This is an advantage because randomization helps to control for order effects among the treatments.

The following example illustrates one of the many applications of the ATD and demonstrates how the design can be modified to investigate different problems. In it, the ATD was used to demonstrate that goldfish can be taught to make a difficult temporal discrimination using operant conditioning techniques.

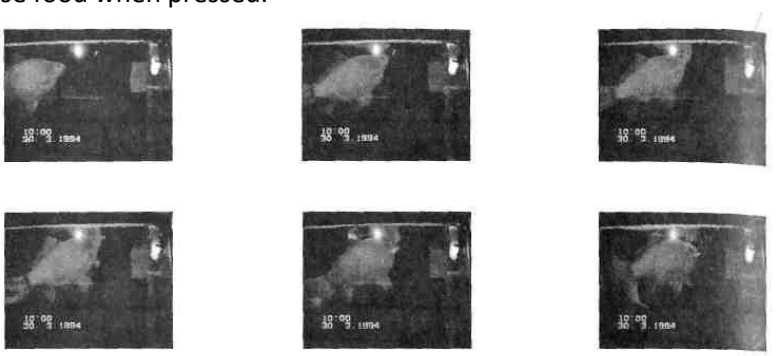
9.4.2Illustration of the ATD Design: Temporal Discrimination in Goldfish

In fixed-interval schedules of reinforcement, particular behaviors are reinforced after a specific time interval, for example, every 10 minutes, every hour, or even every 12 hours. Mammals can coordinate their behavior accurately to long intervals of time and can be taught to start bar pressing close to the start of the reinforcement period, even when there are no external cues for the animals to use to time the interval. These findings suggest the presence in mammals of a "biological clock," a cyclical physiological process that the animals use as a stimulus for marking the passage of time.

Gee, Stephenson, and Wright (1994) studied whether goldfish are able to learn the kind of long-interval temporal discrimination that mammals can. Demonstrating that fish can make such discriminations would be of practical, as well as theoretical, interest. In a technique used in commercial fishing, called "recall ranching," fish are raised in open waters but conditioned using sound to stay in a particular

location for feeding. Then, when they are large enough, the fish are harvested. If fish can learn long-interval temporal discriminations, such temporal conditioning might be substituted for auditory conditioning, saving the cost of expensive sound-generating equipment.

The subjects in the Gee et al. experiment, 8 goldfish *(Carassius auratus)*, were individually housed in controlled aquariums and exposed to an artificial light cycle of 12 hours of bright light (8 a.m. to 8 p.m.) and 12 hours of dim light (8 p.m. to 8 a.m.). Each tank was equipped with a fish-activated lever that could be programmed to release food when pressed.



One of the goldfish from the Gee et al. experiment approaching and pressing the lever to release food.

The experimental plan was eventually to feed the fish, by means of the lever, for only 1 hour a day starting at 2 p.m. There was no cue to this feeding time except the temporal one. If the fish could learn the temporal discrimination, it was expected that they would start pushing the lever in the minutes preceding 2 p.m. in anticipation of feeding. The study was conducted in five phases (see Figure 3):

Baseline: During the first phase, a fish was put on continuous reinforced feeding (CRF); food was released every time the fish pressed the lever. The baseline continued for 14 days, to observe the fish's natural feeding rhythms.

Shaping: During the second phase, continuous reinforcement was discontinued by gradually restricting the time intervals during which lever pressing was reinforced. On the first day, starting at 2 p.m., the feeding interval lasted 12 hours; it was reduced to 10 hours on the second day. Then, on each succeeding day, it was reduced by 2 more hours, so long as

the fish had eaten on the previous day, until the final feeding time was down to 1 hour, still beginning at 2 p.m..

Restricted *Feeding:* The-1-hour-perday feeding interval, beginning at 2 p.m., was continued for 4 weeks.

Extinction*:* During the 6-day extinction period, lever pressing no longer resulted in reinforcement. The fish was not fed for 6 days.

Baseline: Following extinction, the baseline condition of continuous reinforcement (CRF) was reestablished.

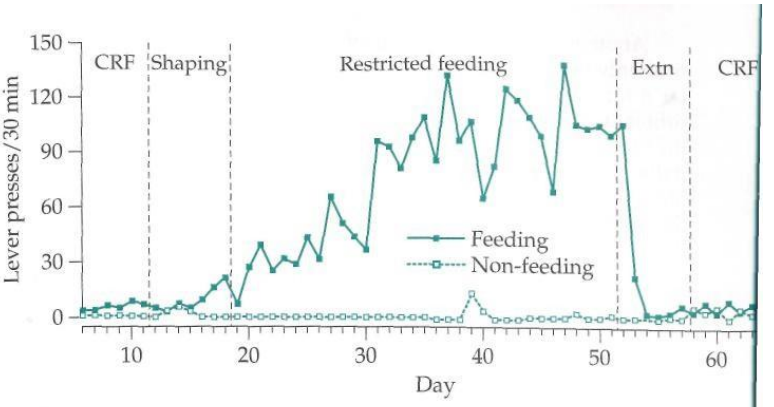


FIGURE 3
Goldfish lever presses during the 30-minute period before feeding and during a 30-mi
control period before no feeding. (From Gee et al., 1994.)

The observations in the study alternated between two time periods, from 1:30 p.m. to 2 p.m., a 30-minute period prior to the 1-hourper-day restricted feeding time; and from 1:30 a.m. to 2 a.m., a 30-minute period selected as a control condition. The feeding and control conditions are the two treatments in this alternating treatment design. The total number of lever presses made by the fish was recorded during each of these periods, throughout the course of the study. Order effects were not an issue because the control condition involved only passive observation of the subject.

The results of the experiment for one fish are shown in Figure 3. Notice that, in contrast to Skinner's cumulative record, this graph charts the number of lever presses made by the fish in two separate 30-minute time intervals over the 60 days of the experiment. The blue squares show the rate of lever pressing in the 30 minutes prior to the 1-hour feeding period; the open squares show the rate of lever pressing for the 30-minute no-feeding control period.

The baseline responding for this fish was low for both time periods. During shaping and the restricted feeding phases, the lever pressing prior to the 2 p.m. feeding time increased, with no change in the rate of response in the control condition. Lever pressing dropped back to the baseline level during extinction, remaining low in the terminal baseline phase. The results provide clear evidence that this fish acquired the temporal discrimination and the results replicated with other fish. One fish died during the experiment, but the six remaining fish behaved similarly to this one.

9.4.3 Simultaneous Replication Designs

ABAB and alternating-treatment sequential replication designs require the experimenter to withdraw a treatment or switch between treatments. Although this does not present an ethical dilemma in many cases, in some it does. If the subject's disorder is serious or long-standing, it may be inadvisable to withdraw an effective treatment or switch to a new one, to make the experiment internally valid.

Another problem with sequential replication designs is that they are practical only when there are minimal carryover effects of the treatment (see Chariter 6 for a discussion of carryover effects). Carryover effects often are not a problem. In Shane's experiment on facilitated communication, for example, the "treatments" only involved presenting pictures to the subject and her facilitator. Because neither the child nor the facilitator was given feedback on the outcome of each trial, the result for one trial couldn't have affected the result for the next trial. But the possibility of carryover effects rules out the use of this design for some problems. When students are taught new skills, like mathematics, writing, or a foreign language, for example, the learning endures beyond the end of the teaching period, making an ABAB design or ATD inappropriate for such cases.

Simultaneous replication designs avoid both the ethical problem of switching between treatments and the problem of carryover effects. In this type of design, two or more O X O subexperiments are carried out at the same time on the same subject. A study with three subexperiments would be diagrammed as:

Subexperiment 1: OXOOOOOOOO

Subexperiment 2: OOOOYOOOOO

Subexperiment 3: OOOOOOOZOO

where X, Y, and Z denote the beginnings of different treatments.

The subexperiments all are designed to test the same type of treatment, but they differ either (1) in the behaviors being treated or (2) in the settings in which the treatment is applied. A child might be given a behavioral treatment to reduce three different inappropriate classroom behaviors, like talking to other students, walking around the room, and not attending to work, for example, or the child might be treated for the same problem behavior, say, walking around the room, in three different classrooms.

This type of design is called a *multiple baseline design* because the design has two or more distinct baselines, one for each subexperiment. The study begins by making observations on all the baselines. After recording the behaviors for a period of time, a treatment then is started for one subexperiment. Then, after a second interval of time, another treatment is begun, and so on, until all the treatments have been introduced. A treatment, once started, continues throughout the study.

Each subexperiment is a replication, a separate test of the effectiveness of the treatment. The treatments begin at different times to guard against confounding by an external event. If the behavior in each separate experiment responds to the treatment, the evidence of its effectiveness is clear. The continuous observation of multiple behaviors or of the same behaviors in different contexts also provides a check on maturation.

The following experiment used the multiple baseline design in one of the first experiments to show the successful application of behavioral techniques to reduce depressive behavior in a child.

9.4.4 Illustration of the Multiple Baseline Design: Behavioral Treatment of Depressive Behaviors

A 10-year-old boy, Dale, who was diagnosed as having a major depressive disorder, had been admitted to a children's psychiatric intensive care unit because of the seriousness of his symptoms. In addition to other symptoms of depression, Dale was uncommunicative in interpersonal situations. Frame, Matson Sonis, Fialkov, & Kazdin

(1982) used a multiple baseline design to test the effectiveness of a behavioral approach to reducing his depressive behaviors.

The behaviors that were selected for reduction were classified into four types:

1. Inappropriate body position (e.g., turning away from the interviewer, covering his face with his hands, bending backward or forward inappropriately).
2. Lack of eye contact (e.g., not looking the interviewer in the eye while talking).
3. Poor speech quality (e.g., speaking too softly or in a garbled fashion, answering questions with one or two words, waiting more than 3 seconds before answering).
4. Bland affect (e.g., lack of emotional tone in voice, lack of hand gestures while speaking).

The frequency of each of these target behaviors was assessed at the beginning of the experiment by having Dale role-play situations that were likely to come up in his home or in the hospital, for example, helping another child to get food in the cafeteria. Each day Dale was asked to role-play 12 different situations. During these role-plays, each of the four target behaviors was rated as present or absent. Dale was assigned 0 to 12 points for each behavior daily, depending on his performance in the role-plays. The higher the score, the more evidence there was of depressed behavior and the less the evidence of the treatment's effectiveness.

A behavioral skills training program, involving instruction, role-playing, modeling, and performance feedback, was developed to reduce these four target behaviors. The design of the experiment is shown in Figure 4. The study lasted for 28 days, with a one-day follow-up 12 weeks after the end of the study. Baselines were recorded for each behavior for the first 8 days. Then, on day 9, the treatment sessions for body position and eye contact were started; on day 14, the treatment for speech quality began; and, finally, on day 20, the experimenters began treating bland affect. Once started, a treatment continued until the 28th day of the study.

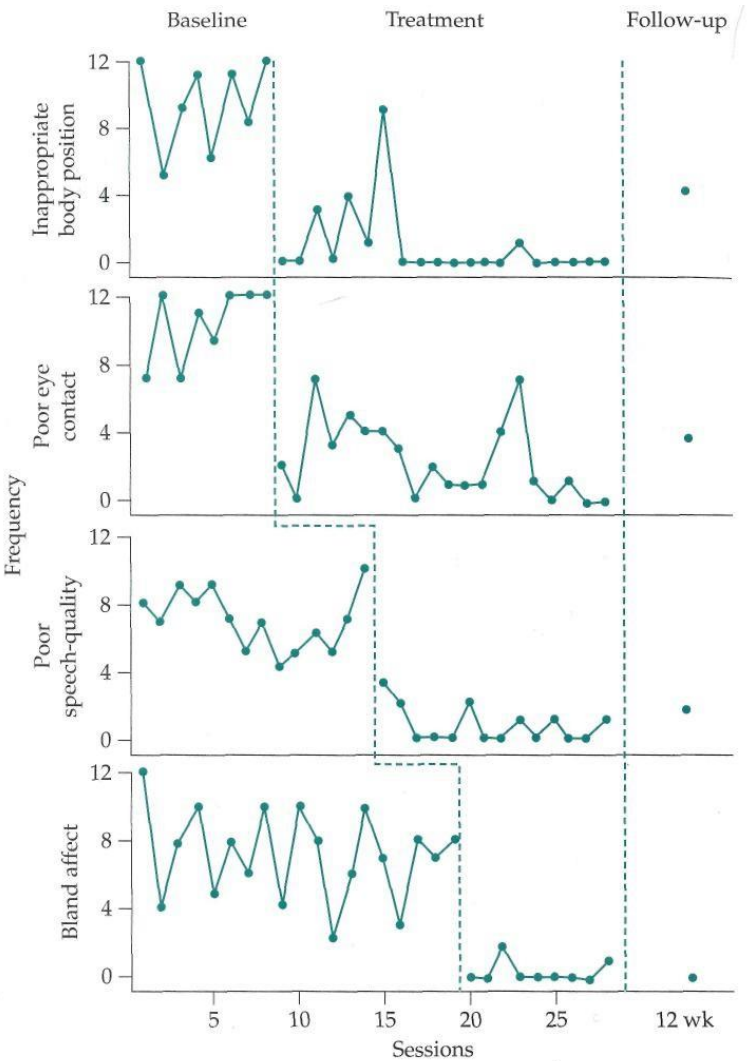


FIGURE 4
Behavioral record for depressed boy during a multiple baseline study. (From Frame et al., 1982.)

Figure 4 shows the frequency of the target behaviors in each phase of the experiment. The results reveal a marked reduction in each of the target behaviors after the treatment for it began and no improvement before that. Bland affect, for example, did not improve when the treatments for eye contact and speech quality were started but did reduce after its direct treatment began. Finally, the treatment effects lasted well beyond the end of the experiment, as indicated by the observations made at the 12-week follow-up.

9.5 Principles Of Design And Analysis

In single-case research, the experimenter has to make choices like those made by researchers using between-subjects designs: The single-case researcher must (1) select a particular experimental design; (2) decide on the size of the study, that is, the number of observations per subject, the number of experimental conditions; (3) determine how to maximize internal and external validity; and (4) make decisions on how to analyze the results.

Statistical procedures have been developed for dealing with some of these design and analysis issues in between-subjects designs: Designs can be compared on their relative power to detect treatment effects; the number of subjects needed to ensure a desired level of power can be calculated; and data analysis can be done using descriptive and inferential statistics. But statistical procedures for single-case studies have not been developed to the same extent as they have for between-subjects designs. So in single-case studies, design and analysis decisions must be based on established practice and rules of thumb rather than on formal principles.

9.5.1 **Choice of Design**

The main design decision in single-case research is whether to replicate sequentially (ABAB designs and ATDs) or simultaneously (multiple baseline designs). The experimenter's choice will depend on the specific treatments being studied. ABAB or alternating treatment designs are practical only when carryover effects are minimal. Ethical concerns also may rule out withdrawal or alternating treatment designs or even the baseline phase of designs. Patients should be given the best known treatment or experimental treatments that promise even better results. Patients with serious problems cannot be switched to placebos, or inferior treatments, or observed for any length of time without treatment (as in the baseline).

When ABAB designs and ATDs are not possible, multiple baseline designs may be appropriate. But these designs are more complicated to conduct than ABAB studies because they involve multiple measures. They also have a potential limitation: If in a design with, say, three baselines, Ol, 02, and 03, behaviors 02 and 03 should change after the treatment for Ol, then the different experiments would not yield

independent evidence of the effectiveness of the treatment. Such a study would have questionable internal validity because the threat of history could not be ruled out. Only experience with the specific behaviors and treatments to be studied will help researchers to decide when this might be a problem.

9.5.2 The Size of the Study

Of course, the number of subjects in a single-case design is not an issue, but the number of observations per phase, the length of the phases, and the number of phases are. Although no formal procedures for making these decisions have been established, the following guidelines have been suggested.

The baseline should have a sufficient number of observations to give a good measure of error. The number of observations needed will depend on the stability of the behavior being observed. If the behavior varies only slightly across several observations, relatively few observations will suffice. When there is considerable variability in the behavior, more observations will be needed. Instead of proceeding in such cases, the experimenter may want to consider canceling the study, trying to figure out the source of the variability, and continuing once the behavior is stabilized. If the baseline shows a pattern or systematic trend, like a steady increase or decrease, or if it has a cyclical form, the experimenter should try to identify the reasons for this pattern before using the baseline to evaluate the treatment.

After the experimenter has established a length for the baseline, the other phases of the study should be matched for duration and for the number of observations collected. Matching the phases for length improves the study by providing a common time interval for the influence of any uncontrolled variables that may affect behavior in the different phases, for example, fatigue, maturation, or daily or weekly cycles. An exception to this guideline would be experiments assessing the effectiveness of treatments for problem behaviors; in such research, the length of time that successful treatments are withdrawn should be minimized for ethical reasons.

The number of phases in a study will depend on the number of treatments being evaluated and the number of replications desired for each subject. A general rule, based on Mill's method of difference (see

Chapter 3), would be to vary only one aspect of the treatment from phase to phase of the experiment. According to Mill, only one antecedent should be varied, while all others are held constant. Consistent with this guideline, the more treatments are being evaluated, the more phases will be needed.

There are no guidelines for deciding on the number of replications in sequential replication designs. Withdrawal studies can involve as few as one replication, as in the standard ABAB study. ATDs, like the design in the goldfish study, use numerous replications of the AB pattern.

For multiple baseline designs, single-case methodologists recommend four separate baselines. If the behavior responds selectively and consistently to four separate treatments, they consider the evidence good that the treatment worked for that subject.

9.5.3 Handling Threats to Internal and External Validity

Single-case designs are vulnerable to the threats to internal validity of history, maturation, instrumentation, and testing (see Chapter 3). Campbell and Stanley (1963) consider the major threat to the O X O design, the building block of all the single-case designs, to be history. As you recall, the threat of history is that external events that occur coincidentally with the application of the treatment are the real causes of changes in the subject's behavior.

In laboratory research, the threat of history can be controlled by isolation. Pavlov, as we noted before, did his research in a "tower of silence" that was isolated from interference by thick sod walls and a moat! In field research, where isolating subjects is not possible, the effects of history can be assessed by replication, using either an ABAB type design or a multiple baseline design. When the effect of the treatment is demonstrated many times with the same subject, the likelihood that a correlated external event produced the effects is negligible. Replication makes the single-case design a powerful procedure.

There is one threat that cannot be controlled by replication, however. This threat is associated with the fact that human participants in such experiments are aware that they are being tested. In single-case research, the same person is observed in all conditions of the experiment and may know when different phases begin. As a result,

the confounding effects of suggestion are especially likely in such studies (see Chapter 12, Planning the Study).

External validity is a more serious problem than internal validity for single-case research. It is this limitation that leads advocates of single-case research to emphasize the importance of replicating findings. As we have seen, replications often are included in the original studies. (The goldfish study, for example, replicated its results with six additional subjects.) But the problem of generalizing results is not restricted to single-case studies. Even when a study is done with, say, 60 subjects, we still do not know if its results would generalize to subjects other than those tested in the research.

For this reason, establishing the generality of findings necessarily proceeds in a hit-or-miss fashion in experimental research. When one investigator publishes a study, others attempt to replicate and extend it. If studies conducted with very different subjects yield a consistent finding, the result is considered to be general. *Meta-analysis* is a recently developed formal method for making sense of the results collected in multiple studies on a particular problem (see Rosenthal, 1984).

9.5.4 Data Analysis

Comparisons between different phases of a single-case experiment often are done by "visual" analysis. The experimenter inspects the graph of the results; if the data show unequivocal results, like most of the experiments discussed in this chapter, the conclusion is straightforward. However, borderline results, that is, results from which different researchers might draw different conclusions, are inevitable.

Researchers disagree on how to proceed when faced with such results. Some experts believe that in such cases the study should be rethought and redone with better controls or a revised treatment. Others think that appropriate statistical analyses are needed to reach an objective conclusion and that such statistics should be used as routinely in single-case experiments as they are in between-subjects research.

If the history of the development of statistics for between-subjects designs is any guide, in the future we can expect to see new statistical methods and more frequent applications of existing statistical

methods for single-case designs. But not all researchers would welcome these methods. Some researchers think that an overemphasis on statistics actually would slow the progress of psychology as a science by shifting the focus away from observed behavior and onto mathematical models. Most likely, in the years to come, we will see advances in both statistical and nonstatistical approaches.

9.6 KEY TERMS

Facilitated communication

Single-case (n — 1) experiments

O X O design

Control group design, with random assignment of subjects

Skinner box

Kymograph

Cumulative record

Baseline vs. treatment phase

ABAB design

Sequential replication design

Simultaneous replication design

ABABCBC design

Changing criterion design

Behavioral criterion

ATD

Block randomization

Multiple baseline design

Meta-analysis

9.7 KEY PEOPLE

Howard Shane

Donald T. Campbell and Julian C.

Stanley

Hermann Ebbinghaus Ivan Pavlov

John Watson

B. F. Skinner

Murray Sidman

Philip Gee et al,

C. Frame et al

9.8 REVIEW QUESTIONS

1. What characteristics of the simple chemistry experiment described m the text make its results so convincing?

2. Explain the notation system used by Campbell and Stanley to experiments.

3. Describe the single-case experiments of Ebbinghaus and Pavlov.

4. What problems can be studied with the control group design that cannot be studied with the single-case O X O design?

5. Diagram a two group control group design with random assignment of subjects to groups.

6. What are the three major limitations of control group designs?

7. How can summary statistics mask patterns that occur in all subjects? Give an example.

8. In which of Skinner's projects was proof at the .05 level of significance irrelevant?

9. What was Skinner's background before studying psychology in graduate school?

10. Describe the three basic phases of a typical experiment Skinner would conduct?

11. What two basic modifications of the O X O design did Skinner introduce? What purposes did these modifications serve?

12. What are the distinctive features of a modern single-case experiment?

13. Describe the two types of replication in modern single-case experiments.

14. Describe how the ABABCBC design might be used to evaluate a new drug,

15. What are the two major advantages of the ATD design over the ABAB design?

16. Describe the ATD design that was used to demonstrate temporal discrimination in goldfish.

17. What problems of an ABAB or ATD design are avoided by using a multiple baseline design?

18. Describe the multiple baseline design that Frame et al. used to evaluate a behavioral treatment for depressive behaviors.

19. What is the main decision researchers must make in designing a single case experiment?

20. What factors should researchers consider in deciding on the length of the baseline?

21. What is the major threat to the internal validity of a single-case experiment and how is it controlled in practice?