CHAPTER 10 FIELD RESEARCH

> The person who must have certitude, who cannot embrace conclusions tentatively, should not be engaged in social scientific research. Norval Glenn

Laboratory experiments show that watching television has an immediate impact on behavior. In one experiment, for example, children who watched a film showing aggressive behavior engaged in more aggressive acts than children who had just seen a neutral control film (Bandura, Ross, & Ross, 1963). In another, children's preferences for sex-neutral toys were affected by seeing a film that presented the toys as appropriate only for a specific sex (Cobb, Stevens-Long, & Goldstein, 1982). But restrictions on the conditions to which people can be exposed and on the types of measures that can be made limit the usefulness of laboratory experiments for discovering how television affects people's lives.

We now know, for example, that television has changed how children spend their time. Before the 1950s, children grew up without television; today, they spend about the same amount of time watching television as they spend in school. On the average, children watch television 3 to 4 hours a day, 7 days a week (Eron, Gentry, & Schlegal, 1994). One major consequence of such extensive television viewing is the displacement of other activities. If people watch television for many hours each week, they can't do other things, like spending this time in community activities or sports. This type of displacement cannot be studied in laboratory experiments.

It also may be difficult to manipulate many variables of interest to psychologists like children's television watching over extended periods of time, in the laboratory. In fact, it may be unethical to deliberately expose children to the kinds of intense and frequent violence found on network television shows.

Clearly, assessing the full impact of television on people's lives requires field research, that is, research done outside the laboratory in the subjects' day-to-day natural environment. Such research conducted over the last 40 years points to the conclusion that television viewing does affect children's behavior and that often the effects are negative (Eron et al., 1994).

Other sorts of projects also require field research. Social programs, like Hawaii Healthy Start, in which home visitors provide overburdened parents with help in raising their children, must be evaluated in the environment outside the laboratory. Studying the behavior of animals in their natural habitats is necessary because animals behave differently in the wild than they do in captivity. Consumer products also often are best evaluated in the field. For example, *Consumer Reports* magazine (1996) tested motor oils using a fleet of New York City taxicabs. The researchers chose this setting because cabs travel many miles, in a short time, under very demanding conditions. *Consumer Reports* (1995) also evaluated psychotherapies by asking readers to report their experiences with different therapies. Surveys such as this are not hampered by the artificial restrictions that are usual in experimental evaluations of psychotherapy (Seligman, 1995).

Finally, some behaviors develop only in natural settings. For example, aggression most often is directed toward family members, friends, and acquaintances. Because such relationships are difficult or impossible to replicate in the laboratory, studies conducted there have had to focus on

aggression against strangers and inanimate objects, different targets than are standard in everyday life.

Although field studies avoid the limitations of laboratory research, this advantage often is purchased with a loss of control. The two hallmarks of laboratory experiments, controlling external conditions by holding them constant and randomly assigning subjects to groups, frequently are impossible in the field, so field researchers have invented other strategies for reducing error. This chapter focuses on how this is done (1) in experimental designs, when random assignment is not possible; (2) in naturalistic observation; and (3) in survey research.

THE DESIGN OF FIELD EXPERIMENTS (QUASI-EXPERIMENTS)

When scientists can control who gets what treatment, as well as when and how behaviors are observed, field experiments are equivalent to laboratory experiments. The real design challenges of field experiments occur when such controls are impossible. Then the task becomes one of developing the best possible research plan, the design that will eliminate the greatest number of potential alternative explanations of the results.

Campbell and Stanley (1963) introduced the distinction between *true* and *quasi-experiments* to distinguish between studies with full experimental control and those falling short of this ideal.

True experiments have random assignment of subjects to conditions and no major threats to internal validity (see Chapter 3).

Fisher's randomized experiments (see Chapter 6) and factorial designs (see Chapter 8) are examples of true experiments.

Quasi-experiments are experiments with threats to internal validity inherent in their designs.

Most of the quasi-experiments that Campbell and Stanley discuss lack random assignment of subjects to conditions and those which have it suffer from the threats of history and/or maturation.

We will use two studies to illustrate how researchers reduce error in quasi-experiments. The first is a study of the effects of introducing television into a community; the second is an evaluation of a program to reduce highway traffic deaths. The highway study, a classic in the literature on field research, was analyzed by Campbell and Ross (1968). For both cases, we present an experimental design that would have been barely adequate for answering the questions of interest to the researchers, followed by the improved design that the researchers actually used.

Pretest-Posttest One-Group Design: OXO

In 1973, the residents of a rural town in Canada, population 658, watched television at home for the first time. Unlike its neighbors, "Notel," as the town was dubbed by the researchers, was in a valley and required a special transmitter to get adequate reception. Many families bought sets in anticipation of the promised transmitter. A group of psychologists, led by Tannis Williams (1986), saw the abrupt introduction of television into this community as a marvelous opportunity to assess the impact of television on people's lives. Because they knew that television would be coming a year ahead of time, they had ample opportunity to plan their study. The research assessed

the effects of television on a variety of behaviors, including reading, vocabulary, creativity, gender-role attitudes, aggression, and leisure activities. We will discuss only the study designed to evaluate television's impact on children's aggression (Joy, Kimball, & Zabrack, 1986).

The study's basic design was a *pretest-posttest one-group design*. Using Campbell and Stanley's notation, which we introduced in Chapter 9, this study would be diagrammed as:

OXO,

where X stands for the introduction of television, the treatment, and O is an observation of aggressive behavior.

Here we see how the OXO design, which we introduced in the last chapter as a basic n = 1 design, can be used in research on groups of subjects. The two observations can be made on different groups of people, in a *cross-sectional design,* or on the same people, in a *longitudinal design*. The Notel researchers used a longitudinal design. They wanted to observe the children's aggressive behavior on the school playground before (the pretest) and two years after the introduction of television (the posttest). They believed that the two year interval would be adequate for the effects of television to become apparent.

The design of the Connecticut study to reduce highway deaths also was OXO. The study compared the incidence of accidental deaths before and after an intense program of ticketing speeders and imposing stiff penalties on those who were convicted. This crackdown on speeders followed a year in which a record number of people, 324, died in automobile accidents. The year of the crackdown, fewer people, 284, died in accidents. Because the study focused on the incidence of accidental deaths, the design had to be cross-sectional.

There are serious threats to the internal validity of both of these OXO studies, as described so far. If the children in Notel were found to be more aggressive after two years of television, for example, this might simply reflect the fact that older children (they now are two years older) behave more aggressively than younger ones (the threat of maturation). Or there might have been a change in the financial status of town residents during the two years of the study. The children's aggression might have increased because of their own or their parents' frustrations over their economic circumstances (the threat of history). A reduction in accidental deaths following the Connecticut crackdown might have been due to improved safety standards for new cars, increased seat belt use, or less driving occasioned by a gas shortage (all threats of history). If effects were found in these studies, they also might have resulted from normal year-to-year fluctuations in the measures; Cook and Campbell (1979) call this type of fluctuation *the threat of instability*.

Because the scientists who conducted these studies understood the threats inherent in the OXO design, they planned their research so that they could evaluate them by means of two general methods:

- *Replicating* the design using different measures and subgroups of subjects, and making repeated observations before and after the treatment.
- Using *nonequivalent control groups,* that is, different groups of subjects selected to enable the researchers to evaluate rival hypotheses.

Replicating the O X O Design with Different Measures and

Subgroups

The Notel researchers wanted to find out whether the effects of television on aggression were general, as expected, or limited to one gender or the other, to a certain age group (1st- and 2nd-graders vs. 4th- and 5th-graders), to a particular type of aggression (physical vs. verbal), or to a particular measure of aggression (direct observation of playground behavior vs. peer and teacher ratings). The researchers replicated the *O X O* design separately for each of these variations in subjects and measures to assess the generality of television's effects.

They found that television was associated with increased aggression on each measure, for both genders, and for all age groups. The consistency of these results argued against explaining them as normal year-to-year fluctuations because such fluctuations would be unlikely to affect different subgroups and measures uniformly.

To evaluate the rival hypothesis of maturation (that older children might be more aggressive), the experimenters repeated the study using cross-sectional data. They compared first graders' aggression before television to the aggression of first graders two years later, after television was introduced. Because the comparison groups were the same age, maturation could not explain observed differences in their levels of aggression. As expected, the researchers found that the children who had watched television were more aggressive.

The researchers used nonequivalent control groups to test for one additional major threat in the Notel study, history, the possibility that event(s) other than the introduction of television might have produced the increase in aggression. We will discuss their strategy for evaluating the effects of history later in the chapter.

Replicating the OXO Design with Repeated Measures, Time-Series Designs

The researchers studying the Connecticut crackdown did not replicate their OXO design with subgroups of drivers or multiple measures. Instead, they decided to replicate the pre- and posttreatment observations using a different design. This *interrupted time-series design* is diagrammed as:

FIGURE 1

Driving fatalities in Connecticut, 1951-1959. (From Campbell & Ross, 1968.)

○○○○○○○**X**○○○○○○○

The addition of several pretreatment observations in this design allowed the researchers to measure the natural variability in traffic fatalities and to detect systematic trends (upward or downward changes) in them. These repeated observations provided a benchmark for evaluating changes in trend following the treatment. In their absence, it is difficult to judge whether a change immediately following the treatment results from it or from normal fluctuation. You will recall that this strategy also is used in $n = 1$ research; in such designs, the repeated pretreatment Os establish the baseline for evaluating the effects of the treatment on the subject (see Chapter 9).

In the Connecticut study, each O refers to an observation of automobile fatalities in a given year. The data for the 5 years prior to and the 4 years after

the crackdown are shown in Figure 1.

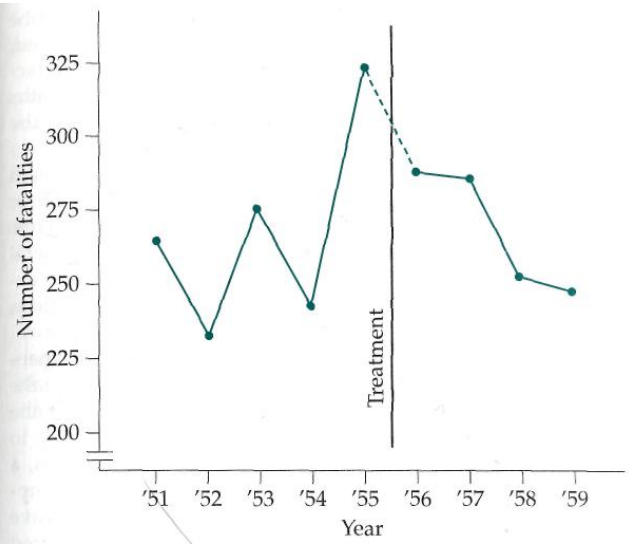


Figure 1 illustrates an upward trend in fatalities prior to the crackdown and a downward trend after, showing a reduction in deaths each year after the new policy. Before the crackdown, there were two increases, a smaller increase from 1952 to 1953, and a larger increase from 1954 to 1955. The large increase just prior to the treatment is a good example of the kind of problem encountered in field research. This alarming increase was what led Connecticut officials to initiate the program; it also made evaluation of the program's effect difficult.

Any year following one with such a large increase in fatalities would be likely to show a reduction in fatalities, even when no treatment is introduced. This is because the increase probably would be caused by a coincidence of factors that would be unlikely to repeat two years in a row. The drop in deaths after the crackdown could have resulted from a *statistical regression* to the mean rather than from the program itself.

As you recall, Francis Galton discovered statistical regression in his research on heredity (see Chapter 5). He found that the heights of children of very tall (or short) parents are closer to the population average for height than their parents. But such *regression to the mean* is not limited to height; regression is a general phenomenon that can occur in any series of repeated observations. If any extreme value (high or low) in the series is selected as the starting point in a study, the odds are that the next value in the series will be closer to the series' mean.

Statistical regression is a threat to the internal validity of a study whenever the researcher schedules a treatment following high (or low) scores on the dependent variable (as in the Connecticut study), or selects subjects for the study on the basis of their extreme scores. The extreme scores are likely to "regress," that is, change toward the mean in subsequent observations, a change that could be attributed erroneously to the treatment (Cook & Campbell, 1979). The results for the first year of the crackdown were inconclusive because it was impossible to know whether the reduction in fatalities resulted from regression or the program. The steady reduction in

fatalities observed over the next several years, however, supported the researchers' claim that the program was effective.

The interrupted time-series design, which the Connecticut researchers used, is equivalent to the "AB" design, which we discussed in the chapter on $n = 1$ designs (see Chapter 9); the design has a series of observations before the introduction of the treatment and a series of observations after the treatment. Two other $n = 1$ designs also are used with groups of subjects in field research: the *alternating treatment design (ATD)* and the *multiple baseline design*.

The ATD and the ABAB design are called *interrupted time-series designs with multiple replications* in the field research literature. Cook and Campbell (1979) diagrammed such designs as follows:

$$OXO \;\; O\overline{X}O \;\; OXO \;\; O\overline{X}O \;\; OXO \;\; O\overline{X}O \;\; OXO \;\; O\overline{X}O,$$

where $\overline{X}$ is the absence of the X treatment, and the schedule of treatments, that is, the order of applying $\overline{X}$ or X, can alternate or be randomized. $\overline{X}$ and X also can be different treatments, rather than one treatment and its absence.

Neither the Notel nor the Connecticut studies could use the interrupted time series design with multiple replications, but it was ideal for a medical study comparing two treatments for gunshot or stab wounds to the torso (Bickell et al., 1994). The standard emergency treatment for such wounds is to start an intravenous infusion (IV) of fluids immediately. However, Bickell and his coworkers were concerned that it might be harmful to administer an IV before the bleeding could be controlled by surgery. They designed their field experiment to resolve this issue.

Their study, done at the Ben Taub General Hospital in Houston, Texas, compared immediate IV to IV delayed until the patient was in the operating room. Patients who sought treatment on even numbered days of the week received immediate IVs; those who came to the hospital on odd days got delayed IVs. (The normal ethical requirement that patients give informed consent was waived by three different institutional review boards.) The results of the study showed better survival rates with delayed IV (203 of 289 patients, 70%, lived) than with immediate IV (193 of 309 patients, 62%, lived), $p < .05$.

We discussed the $n = 1$ multiple baseline design in Chapter 9. In one application of this design, several different behaviors are observed for one subject (these are the multiple baselines) and treatments are started for each behavior on a staggered time schedule. The *interrupted time series with switching replications* design, the equivalent group design, also uses the strategy of staggering the start of treatments (Cook and Campbell, 1979). The design is diagrammed as:

Group 1 *OOOXOOOOOOO*

Group *2:OOOOOOOXOOO*

In this design, the same behavior is observed for both groups, with Group 1 receiving the treatment, X, first, and Group 2 receiving it later. Group 2 serves as a control for Group 1 because any historical events that affect Group 1 when it receives the treatment are unlikely to affect Group 2 subjects at precisely the moment that they receive the treatment. Similarly, Group 1 serves as a control

for Group 2; when the treatment is introduced for Group 2, the scores in Group 1 are not expected to change. Although we have diagrammed this design with two groups, it actually could be used with any number of groups, each group serving as an independent replication for testing the effect of the treatment.

Nonequivalent Control Groups

The *randomized control groups design,* which is used so effectively in laboratory experiments, controls well for threats of history and maturation (see Chapters 6 and 8). Although this design can be used in some field studies (e.g., different motor oils were randomly assigned to particular taxis in the *Consumer Reports* study), often randomization is not possible. Because this was the case in the Connecticut crackdown and Notel investigations, the researchers in these studies had to find existing groups that would control as well as possible for history and maturation. The design they chose, the *nonequivalent control group design,* is diagrammed as:

Control Group: O O

Experimental Group: O X O

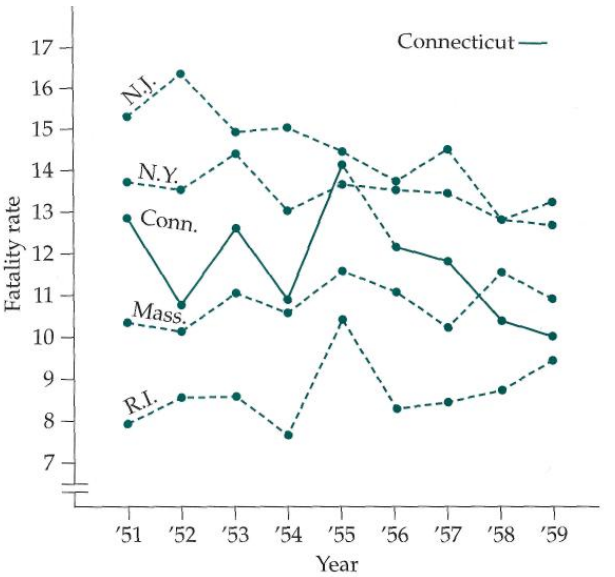


FIGURE 2

Driving fatalities per 100,000 persons in Connecticut and neighboring states. (From Campbell & Ross, 1968.)

The comparable time series design that includes such a control group is called an *interrupted time series with a nonequivalent no-treatment control group time series.* Its diagram is simpler than its name:

Control Group: OOOOOOOOOOOO

Experimental Group: OOOOOO**X**OOOOOO

The control groups in both designs are called nonequivalent because the subjects are not randomly assigned to the groups and, as a result, there may be systematic differences between them. Although such nonequivalence is a problem, a nonrandomized control group usually is preferable to no control group at all.

In the Connecticut study, the best control groups available were drivers in adjacent states. The drivers and driving conditions in these states were similar to Connecticut's, but the adjacent states did not have Connecticut's crackdown on speeders. Figure 2 compares the fatalities in Connecticut and in these control states. Notice that this graph compares fatalities per 100,000 people rather than directly comparing the absolute numbers of fatalities. Expressing the fatalities as a rate controls for differences in the populations of the states. (Although the researchers could have adjusted the rates for differences in the ages and genders of drivers, they did not do so.)

If the pattern of fatalities in the control states was the same as Connecticut's (an upward trend before 1955 and a downward trend afterwards), there would be no evidence of an effect of the crackdown. But this proved not to be the case. Both New Jersey and New York showed continual downward trends before and after the Connecticut crackdown began, and Massachusetts and Rhode Island had upward trends throughout the same time period. Only Connecticut had a change in fatalities coincident with the start of the crackdown providing good evidence for the program's effectiveness in preventing fatal accidents.

The Notel researchers had more difficulty in finding suitable control groups than the Connecticut researchers did. The ideal control for Notel —would have been a twin city that had no television, like Notel, and did not get it when Notel did. But Notel's no-television status was unique in the early 1970s; there was no twin. So the researchers used other small cities adjacent to Notel as controls; these cities had television for several years before Notel got it— "Unitel" had one channel, "Multitel" had several.

The researchers observed residents of these cities and of Notel at the same times to control for history and maturation. Any general effects of history or maturation would be expected to affect the levels of aggression in Unitel and Multitel in the same ways as Notel. Because neither physical nor verbal aggression changed significantly in these control cities during the two years in which they increased in Notel, the researchers rejected history and maturation as explanations of the increased aggression in Notel (see Figure 3).

As the Notel and the Connecticut studies illustrate, nonequivalent control groups play a major role in improving the internal validity of field studies when it is not possible to randomly assign subjects to conditions. Without the evidence from the control groups, both studies would have been inconclusive. Using the control groups, the researchers were able to rule out plausible alternative explanations for the results. In general, the combination of using appropriate nonequivalent control groups and, if possible, taking multiple observations before and after the start of the experimental treatment greatly improves field studies.

NATURALISTIC OBSERVATION

We used the term "observation" in the diagrams above to refer to any measures that are taken on subjects in a study. Actually, the direct observation

of subjects' behaviors is rare in psychological research; ratings, tests, interviews, and questionnaires are much more common. Bakeman and Gottman (1986), experts on the naturalistic observation of social interaction, reported that only about 8% of psychological studies are observational. They implored psychologists to begin observing in their research.

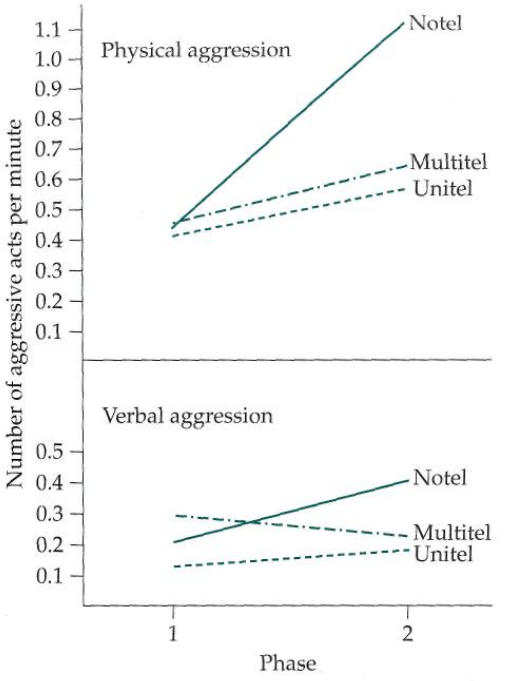


FIGURE 3
Levels of aggression in Notel, Unitel, and Multitel before (phase 1) and after (phase 2) television. (From Joy et al., 1986.)

Bakeman and Gottman admitted that they were looking for colleagues ("we are lonely"), but they also claimed that observation made for good research. They reported that observational studies had permitted them to find consistent results in areas previously recalcitrant to quantitative analysis, including: "how babies learn to interact with their parents and organize their behavior; how young children make friends or are rejected by their peers; how marriages succeed or fail by how spouses interact; how families weather stress or create pathology" (Bakeman & Gottman, 1986, p. 201). They hoped that reporting their successes would encourage others to use observation in their own research.

We already have discussed one study that used observational techniques. The Notel researchers observed children's aggressive behaviors in a natural situation, the school playground (Joy et al., 1986). They did so because they believed that behavioral observations were the best way to determine children's levels of aggression in day-to-day life. The researchers also collected peer and teacher ratings of aggression but concluded that such measures afford less certainty about what children actually do than seeing and recording their activities firsthand.

Direct observations enjoy a validity that can only be hoped for with less direct measures of behavior. To illustrate, Anderson and Lorch (1983, as cited in Joy et al., 1986) compared parents' ratings of the time their children

watched television with records made by a videocamera mounted in the viewing room. The two measures correlated only .21, a low value. Joy et al. suggested that the poor validity of these parents' ratings may help explain why x correlational studies that use them find low or nonsignificant relationships between extent of television watching and other behaviors.

One reason that observations are used less frequently than other methods in research may be that developing a scheme for observing behavior is more difficult than constructing a rating scale or questionnaire. The researcher has to devise a coding system for recording behaviors and observers must be trained to achieve adequate levels of accuracy, a time-consuming process. Goodall (1986) reported that college students were able to make reliable ratings of chimpanzee behavior only after a month of training in the field, and this following extensive instruction on the chimpanzee's ethogram in the classroom.

An *ethogram* is an elaborate, sometimes pictorial, system for classifying the behavior of a species.

Part of the chimpanzee ethogram describing facial expressions is shown in Box 1. Although the term "ethogram" originated in the field of ethology, that is, the scientific study of animal behavior, researchers are beginning to publish ethograms for limited aspects of human behavior. Joan Bottorff and Janice Morse (1994), for example, developed an ethogram for nurse-patient interactions. Whenever anxethogram or more restricted category system is unavailable for a given groupof subjects, which often is the case, researchers must develop their own category systems for recording behavior.

Deciding on Behavioral Units

If you remember from Chapter 3, John Stuart Mill identified two basic steps in research. The first is the event analysis, in which nature, as perceived, is broken down into discrete events, and the second is designing the research, in which events are manipulated or observed to determine cause and effect. Naturalistic observation is concerned with Mill's first step, the event analysis. Because Mill concentrated on his second step, the logic of research design, he presented little information on how to do the event analysis. Today there are many good examples of how to do what Mill called "the event analysis."

The first step in the event analysis is classifying the observed behaviors into discrete categories, so that the behaviors of interest can be differentiated from other behaviors. The major differences between classification schemes can be characterized according to two dimensions: the type of description, *functional versus empirical* (Lehner, 1979), and the *size (or molarity)* of the behavioral unit (Brandt, 1981).

Empirical descriptions classify behaviors in physical terms, for example, muscle movements or bodily movements, such as walking or eating.

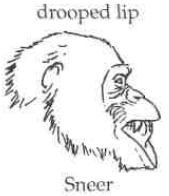
Such descriptions are designed to be objective, avoiding interpretations of behavior. In animal research, they prevent anthropomorphism, the attribution of human characteristics to animals (e.g., describing an animal as "envious" or "disappointed").

BOX 1 CHIMPANZEE FACIAL EXPRESSIONS (FROM GOODALL, 1986)

Relaxed face

Relaxed face with drooped lip

Lip flip



Sneer (fear/threat)

Horizontal pout (distress)

Pout (distress)

Full open grin (fear/excitement)

Compressed-lips face (display)

Low closed grin (fear/excitement)

Full closed grin (fear/excitement)

Full play face

BOX 2 ONE CHIMP'S MORNING (FROM GOODALL, 1986)

Report from the Field

Target: Adult male Jomeo, 14 October 1982

Field Assistants: Esiom Mpongo and Gabo Paulo

0617 Jomeo, Satan, Beethoven, and Freud in nests at KK10.

0634 Climb down. Beethoven presents, pant-grunts to Satan; Satan, hair out, stamps on Beethoven, who screams and falls 3 meters to the ground. Jomeo and Freud pant-grunt, Jomeo climbs a tree and sits. 0640 Satan feeds on *mitati* leaves. Beethoven watches, then feeds on the same. 0655 Jomeo feeds on *mitati* leaves too. They continue feeding and slowly head southeast.

0703 They stop feeding and travel. Freud does not feed.

0708 Freud pant-hoots, all pant-hoot, and Jomeo stamps. These pant-hoots are in response to calls from KK9 and KK8.

0710 The group arrives at KK9 and climbs, feeds on *kirukia* fruit. Goblin, who was feeding at KK9 before the arrival of the group, displays. Beethoven pant-grunts. Jomeo and Satan, no response. Goblin displays around, swaying, hitting saplings.

0715 Goblin climbs, feeds with the others, and food-grunts. Freud vanishes. They stop feeding, climb down. Goblin pant-hoots; the others are silent. All climb, feed on fruits at KK9 with pant-hoots.

0738 They stop feeding, climb down. Goblin displays, silently, not toward anyone. They travel southeast.

Functional descriptions classify behaviors in terms of their purposes or goals, for example, scavenging for food or making a sexual display.

When such goals are obvious (e.g., chimpanzees fishing for termites), functional descriptions replace complicated and tedious empirical accounts (e.g., picks up a stick, walks to the termite mound, inserts the stick into the mound, etc.). Both empirical and functional descriptions are popular and they often are combined in a single descriptive scheme.

In addition, functional and empirical descriptions can use different-sized behavioral units; that is, they can be at either the fine-grained "molecular level" or at the large-scale "molar level."

A description of an animal's movement in terms of the contractions and extensions of individual muscles of its legs is called *molecular;* a description focused on larger behavioral units, like walking or trotting, is *molar*.

Box 2 shows an excerpt from a field record of one chimpanzee's morning (Goodall, 1986). The target chimp is named Jomeo, and his companion chimps are Satan, Beethoven, and Freud. The record gives an empirical description with a relatively large behavioral unit, referring to patterns of behavior, like traveling, feeding, stomping, and "pant-hooting" (a particular cry).

BOX 3 UMTS OF PHYSICAL AGGRESSION (FROM JOY ET AL., 1986)

1. Hits, slaps, punches, or strikes with body part above waist.
2. Hits, slaps, punches, or strikes with a held object.

3. Kicks, steps on, sits on, lies on, or trips with body part below waist.
4. Bites or spits.
5. Pushes, holds, pulls, grabs, drags, or chokes.
6. Snatches property of another (without damage to that property).
7. Damages the property of another.
8. Tries to create a reaction; that is, teases, annoys, or interferes in the activity of another (except where chasing is involved and 11 or 12 is scored).
9. Threatens with some part of the body.
10. Threatens with a held object.
11. Chases another.
12. Chases with a held object.
13. Growls, grimaces, or makes sounds of dislike or anger toward another.
14. Throws or kicks an object at another, except as required (e.g., ball in game).

The observational categories used in the Notel aggression study are listed in Box 3. As you can see there, these are functional categories (e.g., spitting to show aggression) and, like Goodall's, focus on relatively large behavioral units.

The choice of the behavioral unit depends on the purpose of the study. In his research on human facial expressions associated with emotions, for example, Eibl-Eibesfeldt wanted to explore similarities in the expressions of people in different cultures (Eibl-Eibesfeldt, 1967, as cited in Lorenz, 1981). To do this, he recorded their facial expressions on film, which he then played back in slow motion and analyzed for similarities in facial muscle movements. This "microscopic" analysis revealed striking regularities in how culturally diverse groups express emotions like grief and enjoyment.

Reducing Observer Bias

An observer is not a mechanical device that simply records facts. Konrad Lorenz, one of the founders of ethology, expressed this idea when he wrote, "[The observer] is himself a subject, so like the object he is observing that he cannot be truly objective" (Lorenz, 1935; in Lehner, 1979, p. 92). The subjectivity of observers means that their perceptions and interpretations of events are influenced by their knowledge, interests, and needs. Researchers must take pains to minimize biases that they bring to their observations. A good example of the subjectivity of observers came to light in Rubin, Provenzano, & Luria's (1974) study of how parents perceive their newborns. When asked to estimate the height and weight of their babies, parents systematically rated boys as taller and heavier than girls, even though there were no measurable differences between them. Apparently, the parents' beliefs that boys are bigger than girls biased their estimates in the direction of their expectations.

Controlling observer knowledge. Biases can be reduced in some studies by controlling what the observer knows about the experimental situation.

Withholding potentially biasing information from an observer is called *blinding* the observer.

This strategy works because observers can't be biased by what they don't know. The parents' size judgments of their newborns could not have been bi ased by gender if they hadn't known their baby's sex when they made their ratings.

Separating fact from interpretation. However, in many studies, potentially biasing knowledge cannot be hidden; in others, there is no information to withhold, for example, in observations of animals in the wild. In these cases, observers must be trained to separate fact from interpretation and expectation. Goodall's observers, for example, were taught to record behaviors first and then to list possible interpretations of them with a special notation. This procedure reduced bias by constantly reminding the observers to distinguish between fact and interpretation.

Sampling methods. Bias also can be reduced by removing observers' choices about which subjects or behaviors to observe; the less choice, the less the opportunity for personal bias. For example, the observer could be given a set protocol of what to observe. The protocol might specify when the observer should start and stop recording, and which behaviors and subjects to observe during the sampling time interval.

Both Goodall and the Notel researchers used *continuous sampling* focused on one subject at a time. Goodall recorded the behaviors of single chimps from dawn to dusk (Goodall, 1986). In the Notel study, each child on the playground was observed for 1-minute intervals, during which all the aggressive acts of the child were recorded. Once the 1-minute observation was complete for one child, the next was selected, at random, and observed for 1 minute. This protocol continued until each child was observed for 21 minutes spread over two days. The choice of who to observe and for how long was controlled by the researchers, minimizing observer bias.

There are many different behavior-sampling schemes for controlling observer bias. Recording intervals can vary from almost instantaneous to very long periods. Recording can start and stop according to a preset time signal or be contingent on subjects' behaviors. Observations can focus on a single subject or include all the subjects who are visible to the observer. The various sampling schemes that are common in ethology and psychology were analyzed by Jeanne Altmann (1974), who discusses the pros and cons of each method. We recommend her article to anyone planning an observational study.

Observer Reliability

Error also can creep into a study when the rating system is ambiguous, when the raters fail to understand it, and when observers do not pay attention or are bored or tired. These sources of error can be detected by checking the reliability, or consistency, of the observations. In fact, a reliability assessment should be built into every observational study.

To establish reliability, the researcher must demonstrate that an observer's ratings are consistent with those of other raters or of an expert observer.

The *interobserver* (or *interrater) reliability* of observations can be assessed by having different observers rate the same behavior sequence, and then checking the ratings for interobserver consistency.

If the ratings do not agree, this should be a warning to the researcher to retrain the observers or rewrite the rating system. It is wise to establish interobserver reliability before the study begins.

It also is good practice to check on *intraobserver reliability:*

> *Intraobserver reliability* is established by having an observer rate the same behavior sequence at different times throughout the study. Such checks reveal whether observers' ratings are stable (i.e., reliable) or whether they change as the study progresses.

Unreliability can result from fatigue or boredom or because a shift has occurred in the observer's use of the rating system. A lack of intraobserver reliability suggests the necessity of retraining the observer. Additional procedures for assessing reliability are discussed in Chapter 12, Planning the Study.

Reducing Subject Reactivity

In Chapter 1, we presented what may be the first documented example of subject reactivity in psychological research. Benjamin Franklin's commission demonstrated that patients' reactions to Mesmer's magnetic treatment depended upon their *knowing* that they were being treated. Franklin's group eliminated subject reactivity by "magnetizing" subjects without their awareness, a strategy that would be considered unethical by modern standards. Today's researchers use several strategies for handling the problem of reactivity, including: acclimation, concealment, and the use of unobtrusive measures and archival data.

Acclimation and concealment.

> *Acclimation* refers to the tendency of subjects to become accustomed to the presence of an observer with the passage of time.

Konrad Lorenz described acclimation as a methodological ideal for naturalistic research:

> In the observation of highly organized creatures, the methodological ideal is achieved when it has become possible to accustom free-ranging wild animals to the observer to such an extent that their behavior is not influenced by his presence and he can, in fact, conduct experiments with them in a natural environmental setting. (Lorenz, 1981, p. 52)

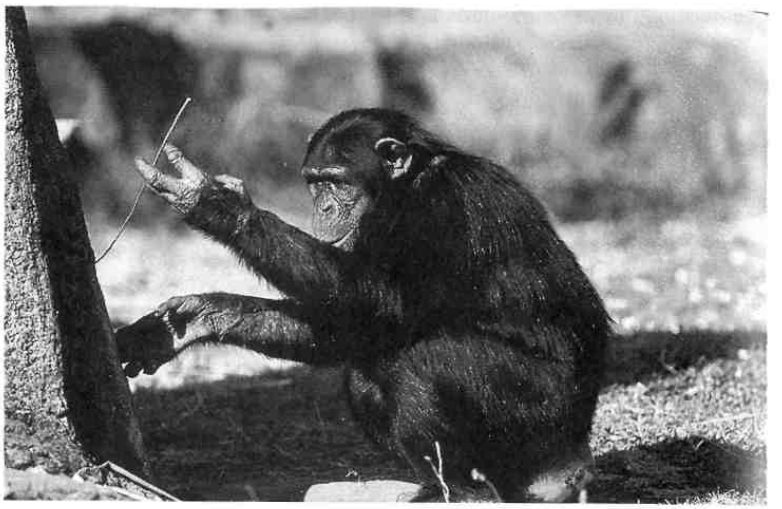
Acclimation occurs both with human and animal subjects. Jane Goodall (1986) was able to learn many new things about how chimpanzees behave in the wild because her subjects gradually grew accustomed to her presence. To promote acclimation, Goodall dressed in similar dull-colored clothes every day and kept at least five meters between her and the chimps. The Notel observers tried to acclimate the children to their presence by coming to the playground for several days prior to rating the children's aggression; while the children were getting used to their presence, the observers took care not to interfere with their play. Although acclimation works well with some animals and with young children, it is less successful with other animals, with older more self-conscious children, and with adults.

Concealment is a good strategy for preventing reactance in animal studies, but there are ethical problems with secretly observing people's private behaviors. One study involving secret observation, in which an observer hid in **a** men's rest room stall and used a periscope to observe how long the men took

to begin urinating, has become a classic example of what not to do in ethical research, for example (Middlemist, Knowles, & Matter, 1976; Koocher, 1977; Middlemist, Knowles, & Matter, 1977).

Concealment also has been used to promote acclimation in studies in which the subjects know they are being observed. Pepler and Craig (1995), for example, described a technology that worked well for audiovisual recording of children's playground behavior. They mounted telephoto cameras in windows overlooking a school playground and put a live microphone and receiver on the particular child they were monitoring. The other children were given dummy microphones and receivers, so they would not know who was being recorded during a play session.

Unobtrusive measures. In cases in which acclimation or concealment would not work, an alternative strategy to reduce subject reactivity would be to use unobtrusive measures of behavior. We are all familiar with the kinds of physical evidence used by forensic experts to reconstruct a crime (e.g., footprints, fingerprints, patterns of blood spatter, DNA from blood samples, and fiber traces). These are called *unobtrusive measures*.



A chimpanzee fishing for termites.

Unobtrusive measures give indirect, circumstantial evidence that certain behaviors have occurred.

Such measures have zero subject reactivity.

Suzuki, Kuroda, and Nishihara (1995) used unobtrusive measures in their study of tool use among chimpanzees in the northern Congo. They were able to get direct eyewitness records of chimps using tools to "fish" for termites on only three occasions, and then only for a moment. As soon as they were aware that they were being observed, the chimps fled into the jungle dropping their tools as they ran (an example of extreme reactivity!).

The researchers retrieved two types of tools: long rigid "perforating" sticks, which they thought the chimps used to poke holes in hard termite mounds, and flexible "fishing pole" sticks with a frayed end, with which they fished for termites. When a chimp sticks the fishing pole into the termite mound, the termite soldiers bite the frayed end and are pulled out by the fish-erchimp.

When the researchers later analyzed fecal samples collected in the area, they found that the chimps were eating soldier termites and that other animals were not, providing indirect evidence of the chimps' tool use. Fecal samples often prove to be a good source of information on the diets and location of animals in the wild.

Unobtrusive measures are less frequent in research on people because more direct ways of observing them usually are available. Nevertheless, unobtrusive measures can prove invaluable in research on people, as they were in one probe into alleged tampering with a standard academic test by school officials in a prestigious elementary school in Connecticut (Avenoso, 1996). No direct eyewitness reports of tampering (changing students' answers to increase scores) were available, but an analysis of the answer sheets for the Iowa Test of Basic Skills produced damaging evidence.

The publisher of the test compared the number of erasures on the answer sheets of the suspected school with the sheets from other schools in the same district. They found that the suspect school had 26.4 erasures per student, compared to only 7 erasures per student at the comparison schools; and 89% of the erasures for the suspect school resulted in correct answers, compared to only *65%* for the comparison schools. In addition, the analysis revealed several different patterns of pencil strokes on many of the answer sheets from the suspect school, suggesting that more than one person had answered questions on the same form. At the time of the newspaper report, the investigation was ongoing.

Archival data. The availability of *archival data* allows researchers to conduct studies that otherwise would\be impossible. Recall, for example, that the researchers in the Connecticut highway study did not make observations of their own, but relied instead on data collected by state agencies. Because we are acclimated to record taking by government and other agencies, subject reactivity is likely to be reduced with such data. Box 4 discusses several archives that can be used in psychological research.

At first thought, the extensive databases and wide availability of archival data seem to be a researcher's dream. The data from elaborate scientific samples of people interviewed by professionals are readily available for your own analyses. In fact, for many projects this truly is a dream come true; but for others, archival data may just result in insomnia.

First of all, the data archived may not answer the precise questions that the researcher is posing. After all, it would be incredible to find that the precise data you need to test your original ideas has already been collected by a state agency! Archival data will not help if original questions or procedures are needed to test a hypothesis.

Second, archival data may be flawed. Webb and his colleagues (Webb, Campbell, Schwartz, & Sechrest, 1966) identified two major problems with archival records, selective deposit and selective survival.

Selective deposit refers to systematic bias in how archival records originally were recorded and in how data were selected for the archive.

Selective survival is the bias introduced when certain types of records in an archive are lost or destroyed.

A study comparing crime rates in Seattle and Vancouver (Sloan et al., 1988), for example, ran into the common problem of selective deposit.

BOX 4 SOURCES OF ARCHIVAL DATA

The federal government is a major source of information on the population of the United States (the census is taken every 10 years), labor and farm information, import-export data, and information on education and safety. State and local governments, private and commercial enterprises, and the United Nations all publish archives.

In addition to such records, there also are large-scale research projects devoted to data collection and storage. Peter Marsden has edited a series of user guides to major social science databases for Sage Publications. The first guide in the series describes the General Social Survey (GSS), a project that began in 1972 (Davis & Smith, 1992). This survey, which is conducted every year, samples English-speaking adults living in households in the United States. Its questions and data, covering such topics as employment, sex, family life, education, religion, politics, crime, health and television viewing, are available to the public. This survey is associated with similar surveys in other countries, permitting cross-cultural research on many topics.

The Henry A. Murray Research Center at Radcliffe College is another major social science data archive. The Murray Center is a repository for data collected by many researchers on a variety of topics. The studies focus primarily on human development, social change, and the life experiences of American women. The center hosts visiting scholars and offers small grants to cover some research expenses incurred while using the archives.

The Aesthetics Research Archive at Boston University is a collection of videotaped interviews with creative artists, including the playwright Arthur Miller and the novelist Saul Bellow, on the creative process. This archive, which was established by Sigmund Koch, the psychologist who edited the six-volume *Psychology: A Study of a Science* (1959-1963), is a major source of data on creativity.

The Internet is rapidly becoming the primary archive for social science databases. "Publishing" archival data on the Internet is easy, and once published the data are available worldwide. Descriptions of the Murray Center archives are available on the Internet, for example, as are statistics collected by the federal government (www. fedstats.gov). If you search the Web for social science archives, you should be able to find many more sites. Commercial organizations also post data on the Internet. The Gallup survey research firm posts the results of recent polls and offers surveys that can be completed at its site (www.gallup.com).

These cities are almost twins, but Vancouver has much stricter gun control laws than Seattle. During the years chosen for the study, the Washington State legislature passed the Domestic Violence Prevention Act, which mandated changes in the reporting of arrests in cases of domestic violence. The result was a marked increase in the number of recorded assaults in Seattle. During this same period, Vancouver had no such change in record keeping. In this case, the researchers avoided bias by not using data collected after the law was passed.

Changes in record keeping are a common occurrence in government archives. New legislation and new technology often mean new record-keeping

systems, and computerization can lead to changes in the records themselves. These problems also can arise in databases for research. The GSS archive, which we described in Box 4, has undergone two changes in its sampling scheme since it began in 1972. Other potential problems in this database, caused by measurement changes, were identified by Smith (1988), who recommends that users check his article to see if their analyses are affected (Davis & Smith, 1992). The Murray Center holds seminars on how to deal with problems in archival data, like incomplete or noncomparable data.

Selective survival is no longer a problem in modern social science databases and government records because records now are preserved on computers. But selective survival can be a serious problem when older records and personal documents are used in research. Controversial or unflattering letters in a private collection may be destroyed and embarrassing political records may disappear, for example.

Webb et al. (1966) recommended the use of multiple measures to overcome potential biases in archival data. If different measures of the same behaviors all point to the same conclusion, confidence in its validity increases.

SURVEY RESEARCH

A critical component of every scientific study is selecting subjects. In this section of the chapter, we discuss how strategies used to select subjects affect the researcher's ability to generalize the results to people not observed in the study, an aspect of external validity.

The two basic procedures for selecting subjects in research are probability and nonprobability sampling.

In *probability sampling,* each subject selected for the study is a member of a larger group of potential subjects, called the *population,* and each member of the population has a probability of being selected that is known and set by the researcher.

An example of probability sampling would be selecting, say, 25 students from a class of 100 by writing the names of all 100 students on separate slips of paper, placing the papers in a bowl, and drawing (as in a lottery) the 25 sample members. Here the population is the group of 100, and the probability of selecting any one person is 25/100 = 1/4.

In *nonprobability sampling,* the population is not specified and the probability of selecting a particular subject is unknown.

An example of nonprobability sampling would be selecting the next 25 people who enter a school cafeteria. The population, the group of potential subjects, is unspecified in this case. A potential subject would be anyone who might enter the cafeteria, and the probability of this event happening is not controlled by the researcher.

Both probability and nonprobability sampling are popular, with virtually all laboratory experiments using nonprobability sampling and all well-done surveys using probability sampling. Probability sampling has both a major advantage and a major disadvantage compared to nonprobability sampling. On the positive side, probability sampling permits findings from the sample to be generalized to the population with a known degree of error. This fact has made probability sampling indispensable for survey research, which has the express purpose of describing a population of subjects. The disadvantage of probability sampling is its expense. If the population is geographically widespread (e.g.,

adults in the United States), the expense of conducting face-to-face interviews or bringing a particular subset of the population to the laboratory may be prohibitive.

Subjects for laboratory experiments in psychology usually are selected by *convenience* (one method of nonprobability sampling). They often are volunteers recruited from introductory psychology classes, the library, or a dormitory. *Quota sampling* is a less frequently used nonprobability sampling strategy for selecting participants.

In *quota sampling,* fixed numbers of specific categories of people (e.g., men and women, first and second borns) are selected for a study.

In clinical studies evaluating therapies for a specific disorder, *typical* or "textbook" *cases* of the disorder usually are selected and patients with dual diagnoses are excluded. *Critical cases* may be selected by case study researchers hoping to discover new insights about a disorder. When potential subjects are difficult to find (e.g., in studies of controlled drugs), the available subjects may be asked to give the names of other potential subjects, the *snowball sampling* procedure.

Such nonprobability sampling schemes share a common problem:

Nonprobability sampling provides no way of determining the accuracy of generalizing results or even of specifying the groups to which the results can be generalized.

Consider a research project on memory that uses college freshmen at an Ivy League college as participants, randomly assigning these student volunteers to the conditions of the experiment. How should the results of such a study be generalized? Because the researchers may consider the processes they are investigating as basic, they may conclude that their results are universal, applying to all adults. But, in fact, there would be no evidence that the results actually would generalize beyond the narrow group of subjects tested in the study.

Probability Sampling

Techniques for probability sampling, the defining characteristic of modern survey research, improve surveys in the same ways that random assignment improves experiments. Both methods avoid systematic biases in selecting (or, in the case of the experiment, assigning) subjects, and both methods provide

TABLE 1 PROBABILITY SAMPLING METHODS

Type	Description
Simple random	Each member of the population has an equal probability of being selected.
Systematic	First subject is selected at random; the rest of the sample is selected from the list of population members at fixed intervals

Stratified	The population is classified into subgroups and simple randomsamples of the desired size are selected from each subgroup.
Cluster	Subjects in the population are classified into clusters, which then are selected at random.
Multistage	The final sample is selected by means of two or more different sampling schemes done in order.

researchers with a way to calculate the potential error introduced by the sampling (or assignment).

The basic probability sampling schemes are listed in Table 1.

In *simple random sampling,* the sample is selected so that each member of the population has an equal probability of being included.

Simple random sampling can be done by using a table of random numbers or a computer programmed to generate random numbers. In simple random sampling, each person in the population is assigned a code number and then numbers are selected using a random numbers table, or the computer generates numbers so that each member of the population has an equal chance of being selected. In random digit dialing, a popular form of simple random sampling, a computer dials telephone numbers selected at random for a certain area code and exchange.

Systematic sampling is an alternative to simple random sampling.

In *systematic sampling,* the sample is selected from the list of population members by randomly selecting the first subject and then selecting the other members of the sample according to a prearranged scheme—by taking, say, every 5th, 10th, or 20th person on the list, until the sample is complete.

To select a sample of 100 from a population of 2,000, the first person would be chosen at random from the first 20 people on the population list. If the 11th person is selected, then the 31st, 51st, 71st, and so on, members also would be included. The advantage of this method over simple random sampling is that it is easier to do, and if the population list is in random order, systematic sampling is equivalent to simple random sampling.

Simple random sampling and systematic sampling give every member of the population an equal chance of being selected for the sample. This feature may not be desirable, however, if researchers want their samples to end up with set numbers of different categories of subjects (e.g., equal numbers of men and women or equal representation of particular age groups). Such control over the size of subgroups can be achieved with *stratified sampling*.

In *stratified sampling,* the population is classified into subgroups and simple random samples of the desired size are taken separately from each subgroup.

In both simple random and stratified sampling, the members of the population are selected one at a time.

In *cluster sampling,* the population is classified into subgroups (or clusters) and whole clusters of subjects are randomly selected.

In a cluster sampling of college dormitory residents, the students might be grouped by dormitory floors and a sample of floors selected by simple random sampling. Every student on a selected floor would be included in the sample. If the study used *multistage sampling,* after the initial selection of floors, a different sampling method would be used to select particular residents of each floor. The second stage might use simple random sampling or cluster sampling, say, clustering by rooms. Research on a large, geographically diverse population might involve several such stages.

Cluster and multistage sampling are done for economy, since they often are much easier to do than simple random sampling. Researchers also select these methods when the goals of the research require them to draw samples from a wide geographic region. But these gains in economy are purchased with a loss of accuracy. Cluster and multistage sampling require more subjects than simple random sampling to achieve the same level of accuracy.

Measuring Error

Perhaps you read the results of a *Newsweek* poll that reported that 49% (±4%) of Americans think that the government is withholding information about UFOs *(Newsweek,* 1996). The 49% figure is the percent of people answering this way *in the sample* of 769 adults responding to *Newsweek*'s poll. The ±4% figure is a measure of the error in estimating the percent of voters in the population (all American adults) who hold this opinion *from the sample.* The error, 4%, is added and subtracted from 49% to give a range of possible values, 45% to 53%, with a set probability of including the population value. This range is called a *confidence interval, CI*. Typically, the size of the confidence interval is set so that the probability of including the population value is 95%; this was the case in the *Newsweek* poll.

The *Newsweek* poll estimated the percent of Americans who believe that the government is withholding information about UFOs. In general, the accuracy of such estimation depends on:

- The size of the sample, *n;* as the sample size increases, the error decreases.
- The size of the population, *N* ; as the population size increases, error increases.
- • The population value (the true value) of the percent being estimated— extreme values, such as 99% or 1%, can be estimated with less error than middle values close to 50%. The value of 50% has the greatest error.

The impact of each of these factors on error is discussed in Box 5.

Modes of Administering the Survey

In 1994, researchers at the University of Chicago published the results of the most comprehensive survey of sexual practices ever done in the United States (Laumann, Gagnon, Michael, & Michaels, 1994). Unlike other well-known sex — surveys (like the Kinsey report, the Janus report, and the Hite report), the Chicago survey used probability sampling, employing the same

sampling methods as the GSS survey, which we discussed in Box 4. The Chicago survey involved intensive interviewing of 3,432 adults.

A critical design decision for these survey researchers was selecting the interviewing strategy. There are three general interviewing modes: *face-to-face interviews,* in which a trained interviewer asks the subject questions in person; *telephone interviews,* in which the interviewing is done by phone; and *self-administrating interviews,* in which the subject completes a paper-and-pencil form with no help from an interviewer. They decided on face-to-face interviews. We will discuss their reasons, because they highlight the general issues that researchers think about in selecting one mode over another.

The Chicago researchers considered the telephone survey first because it is convenient and less expensive than face-to-face interviews; there are no travel expenses and time is not wasted traveling between participants. Simple random sampling, with its high accuracy, also can be done with little effort by randomly dialing telephone numbers. (You could be called even if your telephone number is unlisted, because the computer can randomly pick your number!) Although telephone surveys are limited to people who have telephones, in the United States this includes about 93% of households (Kalton, 1983), resulting in little bias.

The problem with telephone interviews is that the method places restrictions on the length and complexity of the interview. Short simple interviews are perfect for the telephone, but experience shows that 45 minutes is the upper limit for a telephone interview (Laumann et al., 1994). Because the Chicago researchers needed a 90-minute interview, they eliminated the telephone method. In this case, the content of the questions also might have presented a problem, because respondents might be reluctant to give honest answers to intimate questions about their sexual practices over the phone.

Telephone surveys also frequently yield a lower *response rate,* the percent of participants contacted who complete the interview, than other methods. A French telephone sex survey, for example, managed only a 65.5% response rate compared to the 80% rate that the Chicago researchers were able to achieve with face-to-face interviewing (Laumann et al, 1994). (The increasing use of telemarketers may be raising people's resistance to participating in telephone surveys.)

BOX 5 FACTORS AFFECTING THE ERROR OF SIMPLE RANDOM SAMPLES

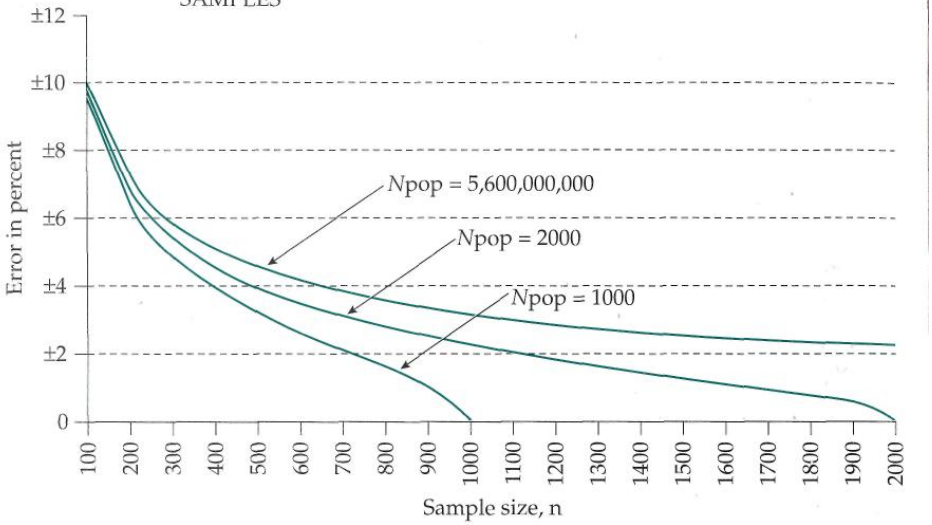


FIGURE 4

Error for simple random samples as a function of the sample size, *n,* and the size of the population, *Npop*.

Figure 4 shows the error in estimating population values from simple random, samples for different sample sizes, *n,* and for three different-sized populations, Npop = 1,000, 2,000, and a whopping 5,600,000,000, the population of the earth. For this graph, the population value was set at 50%, so these errors are the largest possible for the given *n* and *N*.

Notice, first, that when the sample size equals the population size (at $n = N$ = 1,000 and at $n = N$ = 2,000), there is no error. The graph also shows that an increase in the sample size of, say, 100, results in a greater reduction in error when the sample size is small than when it is large. As *n* becomes larger, the reduction in error achieved by increasing the sample size diminishes.

Finally, the graph shows, surprisingly, that the size of the population does not have much effect on accuracy. A sample of n = 500 drawn from a population of 1,000 has an error of ±3.2, but the same sample size drawn from a population of 5.6 billion has an error of only ±4.5. This result is critical for survey research because it means that relatively small samples can give accurate results even for huge populations.

Although simple random samples are used infrequently in survey research, the accuracy of such samples serves as a benchmark for evaluating other sampling schemes. Depending on the categories used in the stratification, stratified samples actually can have less error than simple random samples. Although the error associated with cluster and multistage samples is greater than the error for simple random samples, researchers still may choose these methods, considering their convenience to be worth the loss in accuracy. For example, the GSS database (see Box 4) uses stratified multistage sampling, even though this method requires about 1.5 times the sample size of a simple random sample to achieve the same accuracy (Davis & Smith, 1992)

Self-administered questionnaires also yield low rates of response, lower even than telephone surveys. (How many questionnaires that come in the mail do you answer?) In fact, the problem of response rates led the Chicago researchers to decide against the self-administered questionnaire for their survey. An additional problem with such questionnaires is that they may be too difficult for respondents to answer when the questions involve complicated and confusing *skip patterns* (e.g., answer questions 40-45 only if you answered "yes" to questions 38 and 39). When the interviewer can control the order of such complicated questions, there are fewer mistakes.

The Chicago researchers decided on face-to-face interviews to increase their response rates and because the questions they wanted respondents to answer were complicated. However, they did use self-administered questions for certain sections of their interview. For questions about income and very intimate sexual practices, the interviewer handed the respondents a self-administering form and asked them to put the completed questionnaire in a "privacy envelope." The interviewer never saw the participants' answers to these questions.

The problem with face-to-face interviewing is that it is the most expensive mode of questioning. The Chicago survey, for example, required 220 trained interviewers to conduct 90-minute interviews with 3,432 people. The expense could be justified in this case, however, because the results would contribute to developing public health strategies to fight AIDS.

The Wording of Questions

When well done, probability sampling is not a major source of error in surveys. Samuel Stouffer, one of the pioneers in the scientific use of surveys, discovered in his research that the

> error or bias attributable to sampling and to methods of questionnaire administration were relatively small as compared with other types of variation—especially variation attributable to different ways of wording questions. (Stouffer, 1950, in Payne, 1951, p. 5).

Slight changes in the wording of questions can result in major variations in people's answers, as Stanley Payne (1951) demonstrated in his book *The Art of Asking Questions*. Box 6 presents two of Payne's examples. The first shows the different responses he received to two questions that varied only in whether or not an alternative to the question ("or do you think layoffs are unavoidable?") was stated. Payne split his sample of subjects in half and presented each version to different halves. The results showed that 63% responded "yes" to version one, without the alternative, and only 35% answered "yes" to version two, with it—a 28% difference! In the second example, only one word was changed in the two versions of the question; the word "should" in version one was changed to "might" in version two. This variation resulted in a 19% difference in the percent responding "yes" to the question.

BOX 6 EFFECTS OF WORDING ON THE ANSWERS TO TWO QUESTIONS (FROM PAYNE, 1951)

Question 1:

Version 1: (No alternative is expressed.) Do you think most manufacturing companies that lay off workers during slack periods could arrange things to avoid layoffs and give steady work right through the year?

63% Yes 22% No 15% No opinion

Version 2: (Alternative expressed.) Do you think most manufacturing companies that lay off workers during slack periods could arrange things to avoid layoffs and give steady work right through the year, or do you think layoffs are unavoidable?

35% Yes 41% No 24% No opinion

Question 2:

Version 1: ("should") Do you think anything should be done to make it easier for people to pay doctor or hospital bills? 82% Yes

Version 2: ("might") Do you think anything might be done to make it easier for people to pay doctor or hospital bills? 63% Yes

Differences in people's responses to questions due to wording are much larger than errors due to sampling, and more worrisome. The sampling error can be measured but the effect of a specific wording cannot. In fact, there is no way for a researcher to predict the impact of alternative ways of stating questions. So researchers must rely on their experience, on the results of experiments comparing different wording, and on the advice of experts in writing their questions.

Before developing a questionnaire, it is a good idea to read what the experts have to say (see Payne, 1951; Schuman & Presser, 1981; Converse & Presser, 1986) and gain from their experience. In Chapter 12, we discuss suggestions

from the experts on how to write good questions.

FINAL COMMENTS

This chapter began with the statement by Norbert Glenn, a field researcher, that "people who must have certitude, who cannot embrace conclusions tentatively" should not do social research. This quote highlights the major concern of field researchers, namely, how to deal with the lack of control that usually accompanies studies done outside of the laboratory. Actually, Glenn's advice applies to all researchers, because no research method, laboratory or otherwise, can claim certainty in its conclusions. In fact, the greater control that is possible in laboratory work is accompanied by increased doubt about whether the results will generalize beyond the restricted context of the research.

We can begin to reach certainty in drawing conclusions from research only when the results of many studies, using different methods, both field and laboratory, converge. Research on the health consequences of smoking is a well-known illustration of such convergence. The initial field studies on smoking (epidemiological research) were inconclusive and subject to multiple interpretations. In fact, R. A. Fisher, the developer of modern randomized experimental designs, concluded that the data did not implicate smoking as a cause of disease (Fisher, 1958, as cited in Gould, 1995). However, numerous studies conducted since the original investigations, employing widely different methods and subjects (e.g., experiments on animals and studies on people who have quit smoking) have led to the virtually certain conclusion that smoking causes disease. A similar convergence has been found for the relationship between watching violent television and subsequent aggression toward others (Eron et al., 1994).

Of course, the pace of moving toward certitude can be accelerated by improving individual studies. Such improvements are best achieved by designing and analyzing research results using methodological advances from all branches of science. In this book, we have presented many examples of how methods developed to study one type of problem have been applied with great success to other sorts of problems. Procedures developed for correlational research are used routinely in analyzing experimental data. Experimental designs developed for agriculture now are basic in psychology. Methods developed by ethologists observing animals in the wild now help to reduce error in studies of human interaction in the laboratory. Field studies to evaluate social programs and medical treatments use the $n = 1$ experimental designs invented by the behaviorists.

An interdisciplinary approach to research methodology cannot help but promote the kinds of cross-fertilization that we have described. Methods devised by researchers in other disciplines can provide an endless source of inspiration for innovation in psychology's methods. Researchers in other disciplines, likewise, have much to learn from psychologists.