

First, you may not be in a position to know in advance which variables will be relevant for the matching process. Second, most of the statistics used to analyze the results of experiments assume randomization. Failure to design your experiment that way, then, makes your later use of those statistics less meaningful.

On the other hand, randomization only makes sense if you have a fairly large pool of subjects, so that the laws of probability sampling apply. With only a few subjects, matching would be a better procedure.

Sometimes researchers can combine matching and randomization. When conducting an experiment on the educational enrichment of young adolescents, for example, J. Milton Yinger and his colleagues (1977) needed to assign a large number of students, aged 13 and 14, to several different experimental and control groups to ensure the comparability of students composing each of the groups. They achieved this goal by the following method.

Beginning with a pool of subjects, the researchers first created strata of students nearly identical to one another in terms of some 15 variables. From each of the strata, students were randomly assigned to the different experimental and control groups. In this fashion, the researchers actually improved on conventional randomization. Essentially, they had used a stratified sampling procedure (Chapter 7), except that they had employed far more stratification variables than are typically used in, say, survey sampling.

Thus far I've described the classical experiment—the experimental design that best represents the logic of causal analysis in the laboratory. In practice, however, social researchers use a great variety of experimental designs. Let's look at some now.

Variations on Experimental Design

Donald Campbell and Julian Stanley (1963), in a classic book on research design, describe some 16 different experimental and quasi-experimental designs. This section describes some of these variations to better show the potential for experimentation in social research.

Preexperimental Research Designs

To begin, Campbell and Stanley discuss three “preexperimental” designs, not to recommend them but because they're frequently used in less-than-professional research. These designs are called “preexperimental” to indicate that they do not meet the scientific standards of experimental designs. In the first such design—the *one-shot case study*—the researcher measures a single group of subjects on a dependent variable following the administration of some experimental stimulus. Suppose, for example, that we show the African-American history film mentioned earlier to a group of people and then administer a questionnaire that seems to measure prejudice against African Americans. Suppose further that the answers given to the questionnaire seem to represent a low level of prejudice. We might be tempted to conclude that the film reduced prejudice. Lacking a pretest, however, we can't be sure. Perhaps the questionnaire doesn't really represent a very sensitive measure of prejudice, or perhaps the group we're studying was low in prejudice to begin with. In either case, the film might have made no difference, though our experimental results might have misled us into thinking it did.

The second preexperimental design discussed by Campbell and Stanley adds a pretest for the experimental group but lacks a control group. This design—which the authors call the *one-group pretest-posttest design*—suffers from the possibility that some factor other than the independent variable might cause a change between the pretest and posttest results, such as the assassination of a respected African-American leader. Thus, although we can see that prejudice has been reduced, we can't be sure that it was the film that caused that reduction.

To round out the possibilities for preexperimental designs, Campbell and Stanley point out that some research is based on experimental and control groups but has no pretests. They call this design the *static-group comparison*. For example, we might show the African-American history film to one group and not to another and then measure prejudice in both groups. If the experimental group had less prejudice at the conclusion of the experiment, we might assume the film was responsible. But unless we had randomized our subjects, we would

FIG 1
Thre-

O

i

O

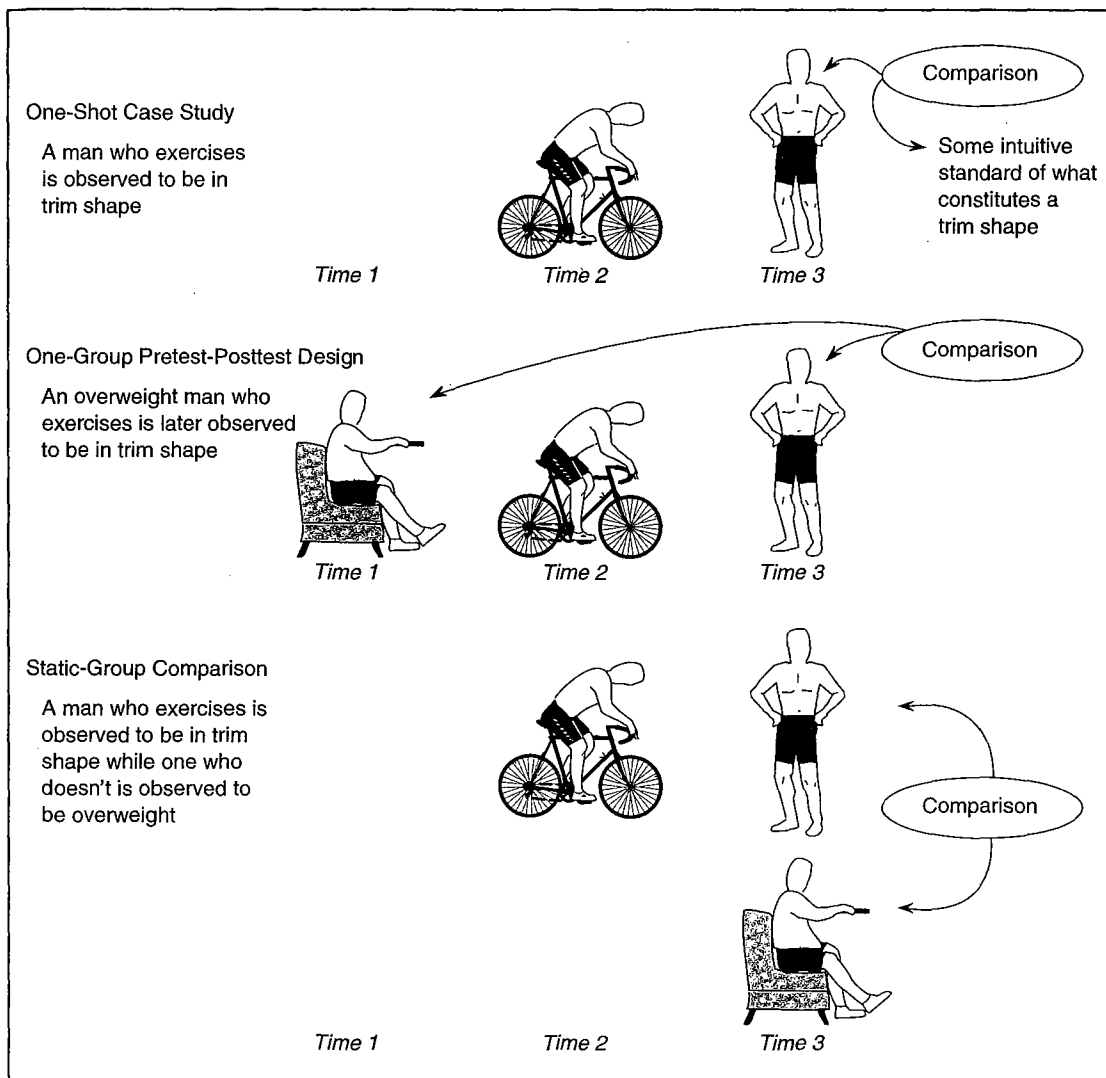
O

S

ha
the
exj

pro
en
rec
fig
thi

FIGURE 8-3
Three Preexperimental Research Designs



have no way of knowing that the two groups had the same degree of prejudice initially; perhaps the experimental group started out with less.

Figure 8-3 graphically illustrates these three preexperimental research designs by using a different research question: Does exercise cause weight reduction? To make the several designs clearer, the figure shows individuals rather than groups, but the same logic pertains to group comparisons. Let's

review the three preexperimental designs in this new example.

The one-shot case study represents a common form of logical reasoning in everyday life. Asked whether exercise causes weight reduction, we may bring to mind an example that would seem to support the proposition: someone who exercises and is thin. There are problems with this reasoning, however. Perhaps the person was thin long before be-

ginning to exercise. Or perhaps he became thin for some other reason, like eating less or getting sick. The observations shown in the diagram do not guard against these other possibilities. Moreover, the observation that the man in the diagram is in trim shape depends on our intuitive idea of what constitutes trim and overweight body shapes. All told, this is very weak evidence for testing the relationship between exercise and weight loss.

The one-group pretest-posttest design offers somewhat better evidence that exercise produces weight loss. Specifically, we have ruled out the possibility that the man was thin before beginning to exercise. However, we still have no assurance that it was his exercising that caused him to lose weight.

Finally, the static-group comparison eliminates the problem of our questionable definition of what constitutes trim or overweight body shapes. In this case, we can compare the shapes of the man who exercises and the one who does not. This design, however, reopens the possibility that the man who exercises was thin to begin with.

Validity Issues in Experimental Research

At this point I want to present in a more systematic way the factors that affect the validity of experimental research. First we'll look at what Campbell and Stanley call the sources of *internal invalidity*, reviewed and expanded in a follow-up book by Thomas Cook and Donald Campbell (1979). Then we'll consider the problem of generalizing experimental results to the "real" world, referred to as *external invalidity*. Having examined these, we'll be in a position to appreciate the advantages of some of the more sophisticated experimental and quasi-experimental designs social science researchers sometimes use.

Sources of Internal Invalidity

The problem of **internal invalidity** refers to the possibility that the conclusions drawn from experimental results may not accurately reflect what has

internal invalidity Refers to the possibility that the conclusions drawn from experimental results may not accurately reflect what went on in the experiment itself.

gone on in the experiment itself. The threat of internal invalidity is present whenever anything other than the experimental stimulus can affect the dependent variable.

Campbell and Stanley (1963:5–6) and Cook and Campbell (1979:51–55) point to several sources of internal invalidity. Here are twelve:

1. *History.* During the course of the experiment, historical events may occur that will confound the experimental results. The assassination of an African-American leader during the course of an experiment on reducing anti-African-American prejudice is one example; the arrest of an African-American leader for some heinous crime, which might increase prejudice, is another.
2. *Maturation.* People are continually growing and changing, and such changes can affect the results of the experiment. In a long-term experiment, the fact that the subjects grow older (and wiser?) may have an effect. In shorter experiments, they may grow tired, sleepy, bored, or hungry, or change in other ways that affect their behavior in the experiment.
3. *Testing.* As we have seen, often the process of testing and retesting influences people's behavior, thereby confounding the experimental results. Suppose we administer a questionnaire to a group as a way of measuring their prejudice. Then we administer an experimental stimulus and remeasure their prejudice. By the time we conduct the posttest, the subjects will probably have become more sensitive to the issue of prejudice and will be more thoughtful in their answers. In fact, they may have figured out that we're trying to find out how prejudiced they are, and, because few people like to appear prejudiced, they may give answers that they think we want or that will make them look good.
4. *Instrumentation.* The process of measurement in pretesting and posttesting brings in some of the issues of conceptualization and operationalization discussed earlier in the book. If we use different measures of the dependent variable in the pretest and posttest (say, different questionnaires about prejudice), how can we be sure they're comparable to each other? Perhaps prejudice will seem to decrease simply because the pretest measure was more sensitive than the posttest measure. Or if the measurements are being made by the experimenters, their