POLITICAL SCIENCE RESEARCH METHODS

SIXTH EDITION

Janet Buttolph Johnson
University of Delaware

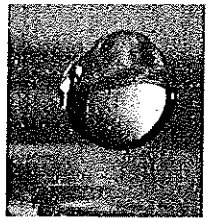
H.T. Reynolds
University of Delaware

with
Jason D. Mycoff
University of Delaware

2008

CQ PRESS

A Division of Congressional Quarterly Inc.
Washington, D.C.



CHAPTER 5

Research Design

CBS News described some political ads aired during the 2004 presidential campaign:

> *In one ad, pictures of Osama bin Laden and Mohammed Atta flash within seconds of an image of John Kerry, as a male voice asks, "Would you trust Kerry up against these fanatic killers?"*
>
> *In another, President Bush is accused of being in financial cahoots with the "corrupt" Saudi royal family and linked to alleged backers of terrorism.*
>
> *"Kerry had a secret meeting with the enemy during the Vietnam War," says a third commercial. President Bush put American soldiers in a "quagmire," another ad charges.*[1]

As we noted in Chapter 1, some political scientists argue that negative advertisements demobilize the electorate.[2] But the CBS network, like many other observers and scholars, believes this approach to campaigning works. They quote Kathleen Hall Jamieson, dean of the Annenberg School for Communication at the University of Pennsylvania: "People are more influenced by negative information than positive information; they are more likely to sway attitudes."[3]

Well, then, what are the effects of negative political advertisements? Do they just cause people to change their evaluations of candidates without affecting their propensity to vote? Do they encourage voting by stimulating interest in campaigns? Or do they simply annoy people to the extent that they want to have little or nothing to do with electoral politics? More specifically, do these types of ads *cause* people to become more or less interested in participating? The discussion in Chapter 1 showed that this is an ongoing and lively issue in political science. It is now time to think about how one might approach a problem of this sort. We require a plan or strategy for collecting and analyzing information in such a way that we can have confidence that our conclusions rest on solid evidence and not on faulty reasoning or mere opinion.

A **research design** is a plan that shows how a researcher intends to study an empirical question. It indicates what specific theory or propositions will be

tested, what units of analysis (for example, people, nations, states, organizations) are appropriate for the tests, what measurements or observations (that is, data) are needed, how all this information will be collected, and which analytical and statistical procedures will be used to examine the data. All the parts of a research design should work to the same end: drawing sound conclusions that are supported by evidence.

A poor research design may produce insignificant and erroneous conclusions, no matter how original and brilliant the investigator's ideas and hypotheses happen to be. In this chapter we discuss various types of designs along with their advantages and disadvantages. As important, we show how a poor research strategy can result in uninformative or misleading results.

Many factors affect the choice of a design. One is the purpose of the investigation. Is it intended to be exploratory, descriptive, or explanatory? Another factor is the practical limitation on how researchers test their hypotheses. Some research designs may be unethical, others impossible to implement for lack of data or sufficient time and money. Researchers frequently must balance what is possible to accomplish against what would ideally be done to investigate a particular hypothesis. Many common research designs entail unfortunate but necessary compromises. As a result, the conclusions that may be drawn from them are more tentative and incomplete than anyone would like.

All research designs to test hypotheses are attempts to (1) establish a relationship between two or more variables, (2) demonstrate that the results are generally true in the real world, (3) reveal whether one phenomenon precedes another in time, and (4) eliminate as many alternative explanations for a phenomenon as possible. In this chapter we explain how various designs allow or do not allow researchers to accomplish these four objectives.

Causal Inferences and Controlled Experiments

Causal versus Spurious Relationships

Let us return to the question of the effects of campaign advertising on voting behavior. A tentative hypothesis is that negative ads, repeated over and over, bore, frustrate, or even anger potential voters and make them think that none of the candidates is worthy of their vote. We might expect that the more citizens are subjected to commercials and advertisements that vilify candidates, the more disinclined they will be to vote. Therefore, turnout will be lower in a campaign flooded with negative ads than in one in which the candidates stick to the issues. We might even be tempted to make the stronger claim that negative political advertising causes a decline in participation.

How could we support such assertions? Just after an election, it might be possible to interview a sample of citizens, ask them if they had heard or been

TABLE 5-1
Voting Intention by Ad Exposure

Y	X	X
Voted?	Yes, exposed	No, not exposed
Yes		100%
No	100%	

Note: Hypothetical data.

aware of attack ads, and then determine whether or not they had voted. We might even find a relationship or connection between exposure and turnout. Let's say, for instance, that all those who report viewing negative commercials tell us that they did *not* vote, whereas all of those who were not aware of these ads did cast ballots. We might summarize the hypothetical results in a simple table. Let X stand for whether or not people saw the campaign ads and Y for whether or not they voted. What this table symbolizes is a relationship or association between X and Y.

This strategy is frequently called opinion or survey research and involves an investigator observing behavior indirectly by asking people questions about what they believe and how they act. Since it does not entail direct observation of their actions, we can only take the respondents' word about whether or not they voted or saw attack ads.

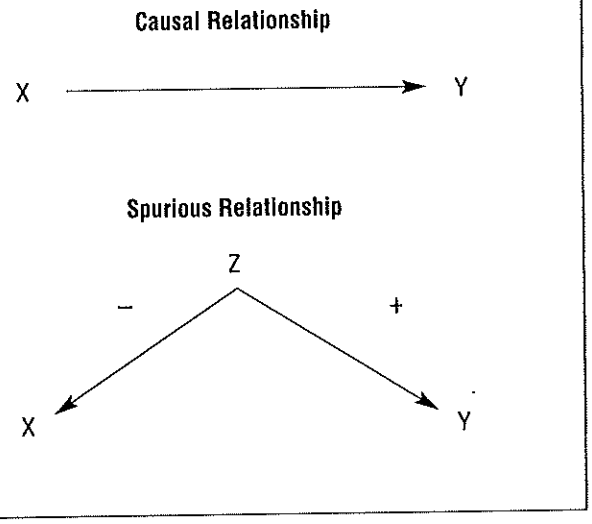
What can we make of these findings? (See Table 5-1.) Yes, there is a relationship. Note that 100 percent of the people who said they were exposed also said they did not vote, and vice versa for those who did not watch any ads. But does that mean that negative advertising causes a decline in turnout? After all, those who missed the ads might differ in other ways from those who saw them. Perhaps they have a higher level of education and that accounts for their higher turnout rate. Or maybe they have a generally strong sense of civic duty and would always vote no matter what the campaigns do or say.

At the same time, people with less education might watch a lot of television and *coincidentally* don't bother voting in any election. If conditions of these sorts hold, we may observe a connection between advertisement exposure and turnout, but it would not be a **causal relationship.** And outlawing negative campaigning would not necessarily have any effect on turnout because the one does not cause the other. In this case the association would be an example of a false or **spurious relationship.**

A spurious relationship arises because two things, such as viewing negative ads and voting, are both affected by some third factor and thus appear to be related. Once this additional factor has been identified and controlled, the original relationship weakens or disappears altogether. To take a trivial example, we might well find a positive relationship between the number of operations in hospitals and the number of patients who die in them. But this doesn't mean that operations cause deaths. Rather, it is probably the case that people with serious illnesses or injuries need operations *and* because of their conditions are prone to die.

Figure 5-1 illustrates causal and spurious relationships.[4]

FIGURE 5-1
Causal and Spurious Relationships

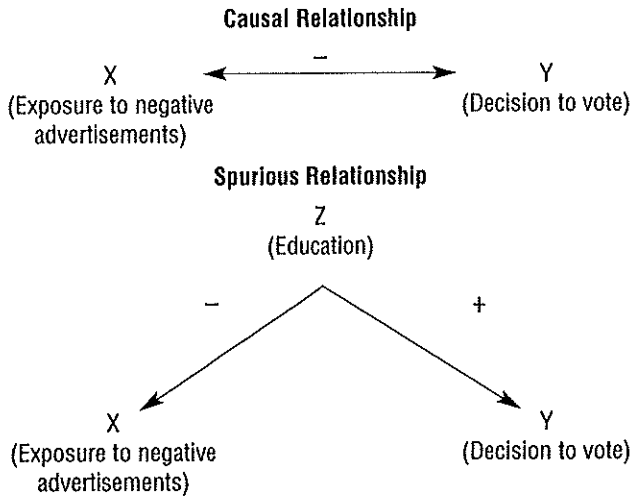


Distinguishing real, causal relationships from spurious ones is an important part of any scientific research. To explain phenomena, we must know how and why two things are connected, not simply that they are associated. Thus one of the major goals in designing research is to come up with a way to make valid causal inferences. Ideally such a design does three things:

1. Covariation. The research must demonstrate that the alleged cause (call it X) does in fact covary with the supposed effect, Y. Our simple study of advertising and voting does this because, as we saw in Table 5-1, viewing negative advertisements is connected to nonvoting, and not viewing the ads is associated with voting. Public opinion polls or surveys can relatively easily identify associations. But to make a causal inference, more is needed.
2. Time order. The research must show that the cause preceded the effect: X must come before Y in time. After all, can an effect appear before its cause? In our survey of citizens, we might reasonably assume that the television ads preceded the decision to vote or not. But however reasonable this assumption might be, we have not demonstrated it empirically. In other observational settings it may be difficult if not impossible to tell if X came before or after Y. Still, even if we can be confident of the time order, we have to demonstrate that a third condition holds.
3. Elimination of possible alternative causes, sometimes termed confounding factors. The research must be conducted in such a way that all possible joint causes of X and Y have been eliminated. To be sure that negative television advertising directly depresses turnout, we need to rule out the possibility that the two are connected by some third factor such as education or interest in politics.

Figure 5-2 shows the possibilities presented by the third requirement. The first diagram (Causal Relationship) shows a true causal connection between

FIGURE 5-2
Models of Advertising Exposure and Voting



X (ad exposure) and Y (voting). The arrow indicates causality: X causes Y. If this is the way the world really is, then attack advertisements have a direct link to nonvoting. The minus sign (–) means the greater the exposure, the less the inclination to vote. (It is called a negative relationship.) The arrowhead indicates the direction of causality, because in this example X causes Y, and not vice versa. In the second diagram (Spurious Relationship), by contrast, the X and Y are not directly related. There is no causal arrow between them. Yet an apparent association is produced by the action of a third factor, Z. The arrow with the negative sign means that people with higher levels of education do not see as many commercials (if any) as those with less schooling. At the same time, citizens with lots of education are likely to vote, whereas those with less education are not as apt to go the polls. (The arrow with the positive sign [+] means education has a positive causal effect on the decision to vote.) Hence, the presence of the third factor, Z (education), creates the impression of a causal relationship between X and Y, but this impression is misleading, because once we take into account the third factor—in language we use later, "once we control for Z"—the original relationship weakens or disappears.

Given the possibility of spuriousness, how do we make valid causal inferences? The answer leads to research design, because how we frame problems and plan their solutions greatly affects the confidence we can have in our results. Asking a group of people about what they have seen and heard in the media and relating their answers to their reported behavior is known in common parlance as *polling*, or opinion research. A more formal term is **survey research**: the direct or indirect collection of information from individuals by asking them questions, having them fill out forms, or other means. (We discuss survey research in Chapter 10.) This approach is perhaps the most common in the social sciences, and it is the one followed in the preceding hypothetical example. A difficulty with survey research, however, is that it is not the best way to make dependable causal inferences. For this purpose many social scientists think laboratory experiments lead to more valid conclusions.

Randomized Controlled Experiments

Experimentation allows the researcher to control exposure to an experimental variable (often called a test factor or an independent variable), the assignment of subjects to different groups, and the observation or measurement of responses and behavior. As we will see, experimental designs theoretically allow researchers to make causal inferences with greater confidence in their dependability than do nonexperimental approaches such as surveys. Although some political scientists do conduct experiments, as Stephen Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino's study of the effects of negative campaign advertising illustrates, most research in the field uses surveys or other methods.[5] The nature of the phenomena of greatest interest to political science, such as who votes in actual elections, requires the use of these methods. Nevertheless, an understanding of experimental design is important because it provides a standard for how to make and evaluate causal inferences and explanations.

As we noted earlier, making a valid causal claim involves showing three things: covariation, time order, and the absence of confounding factors. In theory an experiment can unambiguously accomplish all these objectives. How? Let's look at five basic characteristics of a **classical experimental design**:[6]

1. The experimenter establishes two groups: an **experimental group** (there can be more than one) that receives or is exposed to an experimental treatment, or **test stimulus or factor;** and a **control group,** so named because its subjects do not undergo the experimental manipulation. For example, Ansolabehere and his colleagues had some citizens (the experimental group) watch a negative political ad and others (the control group) watch a nonpolitical commercial. The investigators determined who watched the political ad and who watched the nonpolitical commercial. They did not rely on self-reports of viewership. This control over the two groups is directly analogous to a biologist exposing some laboratory animals to a chemical and exposing others to a placebo (an inactive ingredient or material).
2. Equally important, the researcher randomly assigns individuals to the groups. The subjects do not get to decide which group they join. The random assignment to groups is called **randomization,** and it means that membership is a matter of chance, not self-selection. Moreover, if we start with a pool of subjects, random assignment ensures that at the outset the experimental and control groups are virtually identical in all respects. They will, in other words, contain similar proportions, or

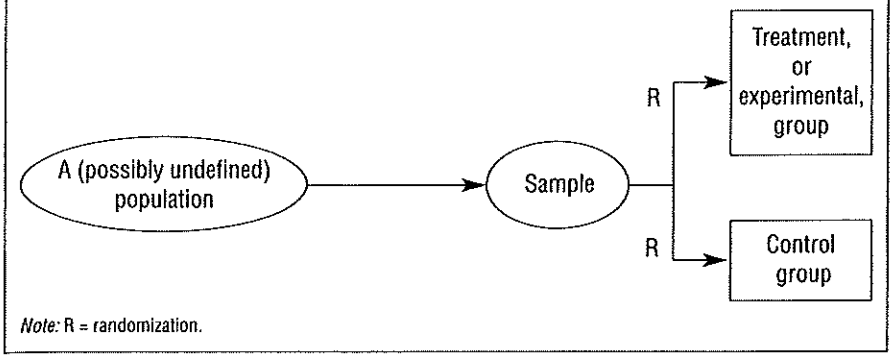
averages, of females and males; blondes, brunettes, and redheads; Republicans and Democrats; political activists and nonvoters; and on and on. On average the groups will not differ in any respect, because they have been created by random placement.* Randomization, as we will see, is what makes experiments such powerful tools for making causal inferences.

3. The researcher controls the administration or introduction of the experimental treatment (the test factor)—that is, the researcher can determine when, where, and under what circumstances the experimental group is exposed to the stimulus.
4. In an experiment, the researcher establishes and measures a dependent variable—the response of interest—both before and after the stimulus is given. The measurements are often called pre- and post-experimental measures, and they indicate whether or not there has been an **experimental effect.** An experimental effect, as the term suggests, reflects differences between the two groups' responses to the test factor.
5. The environment of the experiment—that is, the time, location, and other physical aspects—is under the experimenter's direction. Such control means that he or she can control or exclude **extraneous factors,** or influences, besides the independent variable that might affect the dependent variable. If, for instance, both groups are studied at the same time of day, any differences between the control and experimental subjects cannot be attributed to temporal factors.

To see how these characteristics tie in with the requirements of causal inferences, let us conduct a hypothetical randomized experiment to see if negative political advertising depresses the intention to vote. This case is purely hypothetical, but it resembles the research conducted by Ansolabehere and his associates, and more to the point it shows the inferential power of experiments. (The example also shows some of the weaknesses of this design.)

Our hypothesis states that exposure to negative television advertising will cause people to lose interest in politics and thus to be less inclined to vote. Stated this way, the test factor, or experimental variable, is seeing a negative ad ("yes" or "no"), and the response is the stated intention to vote ("likely" or "not likely"). Now, we recruit from somewhere a pool of subjects and randomly assign them to either an experimental (or treatment) group or a control group. It is crucial that we make the assignments randomly. We do not, for example, want to put mostly females in one group and males in the other because if we find a difference in propensity to vote, how will we be able to

FIGURE 5-3
A Randomized Controlled Experiment



Note: R = randomization.

tell if it arose because of the advertisement or gender? We illustrate the procedure in Figure 5-3.

Note that we draw subjects from some population, perhaps by advertising in a newspaper or giving extra credit in an American government class. This pool of subjects does not constitute a random sample of any population. After all, the subjects volunteered to participate; we did not pick them randomly. But, and here is the key, once we have a pool of individuals we can *then* randomly assign them to the groups. Assume the first subject arrives at the test site. We could flip a coin and, depending on the result, assign that person to the experimental group or to the control section. When the next person comes we flip the coin again, and based on just that result we send the person to one or the other of the groups. If our pool consists of 100 potential subjects, our coin tossing should result in about 50 in each group.

Now suppose we administer a questionnaire to the members of both groups in which we ask about demographic characteristics (for example, age, sex, family income, years of education, place of birth) and about their political beliefs and opinions (party identification, attitude toward gun control, ideology, knowledge of politics). Of course, we would also ask about the main dependent variable, the intention to vote. If we compare the groups' averages on the variables, we should find that they are about the same. The experimental group may consist of 45 percent males, be on average 33.5 years old, and generally (75 percent, say) not care much for liberals. But the control group should also reflect these characteristics and tendencies. There may be only 40 percent males and the average age 35 years, but the differences reflect only chance (or, as we see in Chapter 7, sampling error). Of greatest importance, the proportions on the response variable, the intention to vote in the

next election, should be approximately the same. Thus at the beginning of the experiment we have two nearly identical groups.

After the initial measurement of the variables (the **pre-test**), we start the experiment. So as not to reveal the exact purpose of our research—which if revealed might affect participants' responses to our experiment—we tell the informants that we are interested in television news. Those assigned to the experimental treatment go to Room 101, those in the control panel to Room 106. Both groups now watch identical fifteen-minute news broadcasts. So far both groups have been treated the same. Any differences between them are the result of happenstance.

RANDOMIZATION VERSUS RANDOM SAMPLE
Be sure to remember the difference between *randomization* and a *random sample*. The former, a process, means assigning individuals randomly to groups; it is something an investigator does. A random sample, by contrast, refers to how a small set of subjects is selected from a larger population.

Next, the first set of subjects sees a thirty-second negative ad that we have constructed to be as realistic as possible, while the others see a thirty-second commercial about toothpaste, also as true to life as we can make it. The different treatment constitutes the experimental manipulation (seeing versus not seeing the negative advertisement). After the commercials have aired we show both groups another fifteen-minute news clip. When the broadcast is over we administer (part of) the questionnaire again and measure the propensity to vote. This calculation gives us an indication of the size of the experimental variable's effect, if there is one. Hypothetical results from this experiment are shown in Table 5-2.

Both control and experimental subjects had about the same initial stated intention of voting (68 and 70 percent, respectively), as we would expect, because the participants had been randomized. But the **post-test** measurement shows quite a change for the experimental group—the percentage intending to vote has dropped from 70 to 20 percent, a decline of 50 percentage points. But the control group has changed hardly at all.

So we might conclude that the experimental factor did indeed cause a decline in intention to vote. How can we make this inference? Well, the research design satisfied all the conditions necessary for making such claims. In Table 5-2 we show that the two factors (or variables) covary: those who have seen a negative ad are much less likely to vote than those who did not (20 percent versus 66 percent). We have also established the time order, since we literally determined the timing of the experimental treatment and the subsequent post-test measurement. Finally, and most convincing of all, we have been able to rule out any possible al-

TABLE 5-2
Results of Hypothetical Media Experiment

Group	Before Experiment (Pre-test) Measure of Intention to Vote	After Experiment (Post-test) Measure of Intention to Vote
Experimental	70%	20%
Control	68%	66%

Note: Hypothetical data.

FIGURE 5-4
A Randomized Controlled Experiment: General Framework

	Pre-test		Post-test
R Experimental Group	M_{exp1}	X	M_{exp2}
R Control Group	$M_{control1}$		$M_{control2}$

X = Experimental manipulation

M = Measurements

Experimental effect = $(M_{exp2} - M_{exp1}) - (M_{control2} - M_{control1})$

R = Random assignment of subjects to groups

ternative explanation of the covariation, given that our randomization and experimental manipulation ensured that the groups differed *only* because one received the treatment and the other did not. Since that was the only difference, the gap between the post-test percentages of the two groups could be attributed only to viewing the commercial.

To simplify later explanations, we introduce a general framework for describing research designs, as shown in Figure 5-4. We let X stand for an experimental manipulation (for example, showing a negative campaign commercial) and let the letter M denote the average measurements on the response variable. In the present case M means the percentage of members in a group who said they intend to vote in the next election. The subscripts identify the group and time to which the measure applies. An experimental or treatment effect can be defined in a couple of ways. We are most interested in the change in the experimental group proportions from before and after the test factor was administered compared with those of the control group. Presumably the control group's responses did not change except perhaps by a slight amount due to random errors in collecting and recording the data. Hence, the difference between the post-test and pre-test measures for control subjects should be about zero. But in the experimental group we hypothesized that there will be a noticeable difference in measurements. In this case we would expect that the post-experiment percentage intending to vote will be smaller than the pre-experiment number. We might, then, say that the experimental effect is $E = M_{exp2} - M_{exp1}$, or $E = (M_{exp2} - M_{exp1}) - (M_{control2} - M_{control1})$, which is the same because in this particular case the second term is expected to be zero.

The purpose of an experiment is to isolate and measure the effects of the independent variable on a response. Researchers want to be able to separate the effect of the independent variables from the effects of other factors that might also influence the dependent variable. Control over the random assignment of subjects to experimental and control groups is the key feature of experiments

because it helps them to exclude, rule out, or control for the effects of factors that might create a spurious relationship.

Randomization and the Assignment of Subjects

As we have stressed, the way researchers assign subjects to control and experimental groups is important. The best way is **assignment at random,** under the assumption that extraneous factors will affect all groups equally and thus be canceled out. Random assignment is an especially attractive choice when a researcher is not able to specify possible extraneous factors in advance, or when there are so many that it is not possible to assign subjects to experimental and control groups in a manner that ensures the equal distribution of these factors.

Even if a researcher does assign subjects at random, extraneous factors may not be totally randomly distributed and therefore can affect the outcome of the experiment. This might be true especially if the number of subjects is small. Prudent researchers do not assume that all significant factors are randomly distributed just because they have randomized the study's participants. So, in addition to random assignment, investigators use pre-tests to check to see if the control and experimental groups are, in fact, equivalent with regard to factors known to influence the outcome or suspected of doing so. An especially important measure in this regard is $M_{\text{exp1}} - M_{\text{control1}}$, which should be about zero.

If the researcher knows ahead of time that certain features are related to differences in the dependent variable, he or she can use **precision matching** to control for them. This requires creating matched pairs of subjects who are as similar as possible and assigning one to the experimental group and the other to the control group. Thus no one has to depend solely on randomization to eliminate or control for these factors. One problem with this method is that when many factors are to be controlled, it becomes difficult to match subjects on all relevant characteristics and a larger pool of prospective subjects is required. A second problem is that the researcher may not know ahead of time all extraneous factors. To guard against bias in the assignment of pairs, each member of the matched pairs should be randomly assigned to the control and experimental groups.

One of the biggest obstacles to experimentation in social science research is the inability of researchers to control assignment of subjects to experimental and control groups. This is especially true when public policies are involved. Even though the point of conducting an experiment is to test whether a treatment or program has a beneficial effect, it is often politically difficult to assign subjects to a control group. People assume that the experimental treatment must be beneficial—otherwise the treatment or program would not have been proposed as a response to a public policy problem. Social scientists

nerally lack sufficient authority or incentives to offer subjects in order to t up experimental and control groups.

nterpreting and Generalizing ne Results of an Experiment

ost readers, we hope, have followed the logic of our arguments. But they ust be flabbergasted at the unrealism of the hypothetical example we intro-ced and wonder how anyone could make a definitive statement about nega-ve advertising based on these data, even if we had carried out this experi-ent on real people using real television commercials. Someone might say, his test is invalid." It may be, but before jumping to that conclusion we ed to consider carefully and closely the term *validity.*

nternal Validity

atistical theory tells us that experiments conducted properly can lead to lid inferences about causality. In this context, however, *validity* has a par-cular meaning, namely, that the manipulation of the experimental or inde-endent variable itself, and not some other variable, influenced the dependent ariable. This kind of validity is known as **internal validity,** which means the esearch procedure demonstrated a true cause-and-effect relationship that as not created by spurious factors. Social scientists generally believe that ne type of research design we have been discussing—a randomized con-olled experiment—has strong internal validity. But it is not foolproof.

Several things can affect internal validity. As we have argued, the principal trength of experimental research is that the researcher has enough control ver the environment to make sure that exposure to the experimental stimulus s the only difference between experimental and control groups and that at the utset all comparison groups have the same traits. Sometimes **history,** or vents other than the experimental stimulus that occur between the pre-test nd post-test measurements of the dependent variable, will affect the depen-ent variable. For example, suppose that after being selected and assigned to room, the experimental subjects happen to hear a radio program that under-uts their faith in the electoral process. Such a possibility might arise if a long ag occurred between the first measurement of their attitudes and the start of he experiment. This situation is shown schematically in Figure 5-5. In this in-tance no one would be able to say if the radio program (Z in the figure) or the xperimental treatment produced the effect on voting intentionality.

Another potential confounding influence is **maturation,** or a change in ubjects over time, which might produce differences between experimental nd control groups. To take a different example, subjects may become tired, onfused, distracted, or bored during the course of an experiment. These

FIGURE 5-5
The Effects of "History" on an Experiment

	Pre-test			Post-test
R Experimental Group	M_{exp1}	Z	X	M_{exp2}
R Control Group	$M_{control1}$	Z		$M_{control2}$

X = Experimental manipulation
R = Randomization
Z = Nonexperimental event
M = Measurements

changes may affect their reaction to the test stimulus and introduce an unanticipated effect on post-treatment scores.

The standard experimental research design involves measurement or observation, which is sometimes called testing. However, **testing,** the process of measuring the dependent variable prior to the experimental stimulus, may itself affect the post-treatment scores of subjects. For example, simply asking individuals about politics on a pre-test may alert them to the purposes of the experiment. And that in turn may cause them to behave in unanticipated ways. Similarly, suppose a researcher wanted to see if watching a presidential debate makes viewers better informed than nonviewers. If the researcher measures the political awareness of the experimental and control groups prior to the debate, he or she runs the risk of sensitizing the subjects to certain topics or issues and contributing to a more attentive audience than would otherwise be the case. In this case, we would not know for sure whether any increase in awareness was due to the debate, the pre-test, or a combination of both. Fortunately, some research designs have been developed to separate these various effects.

Selection bias can also lead to problems. Such bias can creep into a study if subjects are picked (intentionally or not) according to some criteria and not randomly. For example, the experimental group might consist of volunteers who differ significantly from nonvolunteers. Sometimes a person might be picked for participation in an experiment because of an extreme measurement (very high or very low) of the dependent variable. Extreme scores may not be stable; when measured again, they may move back toward average scores. Thus changes in the dependent variable may be attributed erroneously to the experimental factor. This problem is called **statistical regression.**[7]

As we stressed, in assigning subjects to experimental and control groups, a researcher hopes that the two groups will be equivalent. If subjects selectively drop out of the study, experimental and control groups that were the same at the start may no longer be equivalent. Thus **experimental mortal-**

ity, or the differential loss of participants from comparison groups, may raise doubts about whether the changes and variation in the dependent variable are due to manipulation of the independent variable. A common selection problem occurs when subjects volunteer to participate in a program. Volunteers may differ significantly from nonvolunteers; they may be more compliant and eager to please, healthier, or more outgoing.

Sometimes, **instrument decay,** or change in the instrument used to measure the dependent variable, occurs during the course of an experiment so that the pre-test and post-test measures are not made in the same way. For example, a researcher may become tired and not take post-test measurements as carefully as pre-test ones. Or different persons with different biases may conduct the pre-test and post-test. Thus changes in the dependent variable may be due to measurement changes, not to the experimental stimulus.

Another possible problem comes from **demand characteristics,** or aspects of the research situation that cause participants to guess the purpose or rationale of the study and adjust their behavior or opinions accordingly. It has been found that people often want to "help" or contribute to an investigator's goals by acting in ways that will support the main hypotheses.[8] Perhaps something about our experiment on political advertising tips off subjects that we, the researchers, expect to find that negative ads depress turnout. The subjects (perhaps unconsciously) may adjust their feelings in order to prove the proposition and hence please us. In this case, the desire to satisfy the researchers' objectives, not the commercials themselves, affects the disposition to vote. This is not a minor issue. You may have heard about double-blind studies in medical research. The goal of this kind of design is to make it impossible for the patients and attendants to identify who is receiving a real experimental medicine and who is receiving a traditional medicine or a placebo.

External Validity

In short, a lot of things can go wrong even in a carefully planned experiment. Nevertheless, experimental research designs are better able to resist threats to internal validity than are other types of research designs. (In fact, they provide an ideal against which other research strategies may be compared.) Moreover, we discuss below some ways to mitigate these potential errors. Yet, even if we devised the most rigorous laboratory experiment possible to test for media effects on political behavior, some readers still might not be convinced that we have found a cause-and-effect relationship that applies to the real world. What they are concerned about without being aware of the term is **external validity,** the extent to which the results of an experiment can be generalized across populations, times, and settings.

One possible objection to experimental results is that the effects may not be found using a different population. Refer back to Figure 5-3, which showed

that participants are selected from some population and then assigned to one of two groups. But the population from which they have been drawn may not reflect any meaningful broader population. Suppose, for instance, we conducted our advertising experiment on sophomores from a particular college. Results might be valid for second-year students attending that school but not for the public at large. Indeed, the conclusions might not apply to other classes at that or any other university. To take another example, findings from an experiment investigating the effects of live television coverage on legislators' behavior in state legislatures with fewer than one hundred members may not be generalizable to larger state legislative bodies or to Congress. In general, if a study population is not representative of a larger population, our ability to generalize about the larger population will be limited.

Another question is whether slightly different experimental treatments will result in similar findings or in findings that are fundamentally different. For example, a small increase in city spending for neighborhood improvements may not result in a more positive attitude of residents toward their neighborhood or the city. A slightly larger increase, however, may have an effect, perhaps because it has resulted in more noticeable improvements.

Threats to the external validity of an experiment also may result if the artificiality of the experimental setting or treatments makes it hard to generalize about findings in more natural settings. In our case, we showed the experimental group just one ad. But in the real world, voters are exposed to dozens and dozens of brief television advertisements at home and elsewhere. They may or may not pay attention to them, or the ads may be "filtered" by comments from family members or friends. None of these conditions was part of our study. Furthermore, as we noted, subjects may react to a stimulus differently when they know they are being studied than when they are in a natural setting.

Despite the difficulty of generalizing the results, experiments are still attractive to researchers because they provide control over the subjects and their exposure to various levels of the experimental or independent variable, and they do permit valid causal inferences, even if of limited generality.[9] Before considering alternative approaches, let us consider another study that used experimentation.

Shanto Iyengar, Mark D. Peters, and Donald R. Kinder attempted to determine the extent to which exposure to televised network news coverage of particular public policy issues can increase the public's awareness and concern about those issues.[10] To test the effect of the hypothesized independent variable (exposure to television news), the researchers used a randomized design. First, they recruited participants for their experiments (paying each participant $20) and had them fill out a questionnaire that included measures of the importance of various national problems (the pre-test). The partici-

pants were then randomly divided into experimental and control groups and, over a four-day period, exposed to videotape recordings of the preceding evening's network newscast.

Unknown to the participants, these newscasts had been edited to include (or exclude) stories from previous newscasts dealing with a particular public policy issue. In one session the experimental group saw newscasts that included stories about alleged weaknesses in U.S. defense capability, while the control group saw newscasts with no defense-related stories. In another session one experimental group again saw newscasts with stories about inadequacies in U.S. defense preparedness, while a second experimental group saw stories about environmental pollution and a third experimental group saw stories about inflation. The day after the last viewing session, participants completed a second questionnaire that again included measures of the importance of various policy issues. All groups except for one reported a significant increase in concern about an issue after they had been exposed to news stories about that issue. The exception involved concern about inflation. In that case, the level of concern about inflation was already so high (a score of 18.5 out of 20) that exposure to the newscasts about inflation had no appreciable effect.

The investigators were sensitive to the measurement **instrument reactivity** and the external validity issues discussed previously. They believed that they had achieved fairly natural viewing conditions (exposure to the news stories took place over several days, in small groups, in an informal setting, without any pressure to pay close attention), that the participants showed no signs of knowing what the experiment was about, and that the participants were fairly representative of a larger adult population. They believed that their experimental results demonstrated that network news coverage could significantly alter the public's sense of the importance of different political issues.

One might argue that the second part of this experiment was not a classical experiment in that there was no control group, only three experimental groups using a pre-test/post-test design. The pre-test/post-test design allows the researcher to compare changes in groups receiving different treatments. Any differences in change may be attributed to differences in treatments. Without a control group, however, change between a pre-test and a post-test does not absolutely establish the factors that caused the change. The researcher can never be sure what would have happened to subjects if no treatment had been given. Nevertheless, there may be good reasons for omitting a control group. Because the researchers had already conducted an experiment that included a control group and that demonstrated the impact of newscasts on issue concern, it was reasonable for them to omit a control group from their second experiment.

Other Versions of Experimental Designs

Now that we have discussed experimental research in general and some problems associated with it, we briefly describe some variations on this approach. Each variation represents a different attempt to retain control over the experimental situation while also dealing with threats to internal and external validity. Although you may not have an opportunity to use these designs, knowledge of them will help you to understand published research and to determine whether the research design used supports the author's conclusions.

Simple Post-test Design

The most basic experiment, the **simple post-test design,** involves two groups and two variables, one independent and one dependent, as before. And subjects are randomly assigned to one or the other of two groups. One group, the experimental group, is exposed to a treatment or stimulus, and the other, the control group, is not or is given a placebo. Then the dependent variable is measured for each group. Using the previous notation, this design may be diagrammed as in Figure 5-6.

Someone using this design can justifiably make causal inferences because he or she can make sure that the treatment occurred prior to measurement of the dependent variable. Furthermore, he or she knows that any difference between the two groups on the measure of the dependent variable may be attributed to the difference in the treatment—in other words, to the introduction of the independent variable—between the groups. Why? This design still requires random assignment of subjects to the experimental and control groups and therefore assumes that extraneous factors have been controlled (that is, were the same for both groups). It also assumes that prior to the application of the experimental stimulus, both groups were equivalent with respect to the dependent variable.

Let us illustrate with a basic example. Suppose we wanted to test the hypothesis that watching a national nominating convention on television makes

FIGURE 5-6
Simple Post-test Experiment

		Post-test
R Experimental Group	X	M_{exp}
R Control Group		$M_{control}$

X = Experimental manipulation

M = Measurements

R = Random assignment of subjects to groups

people better informed politically. Using this research design, we would randomly assign our subjects to a group that will watch a convention or to a group that will not and then measure how well informed the members of the two groups are after the convention is over. Any difference in the level of awareness between the two groups after the convention would be attributed to the effect of watching convention coverage.

The simple post-test experimental design assumes that the random assignment of subjects to the experimental and control groups creates two groups that are equivalent in all significant ways prior to the introduction of the experimental stimulus. If the assignment to the experimental and control groups is truly random, and the size of the two groups is large, this is ordinarily a safe assumption. However, if the assignment to groups is not truly random or the sample size is small, or both, then post-treatment differences between the two groups may be the result of pre-treatment differences and not the result of the independent variable. Because it is impossible with this design to tell how much of the post-treatment difference is simply a reflection of pre-treatment differences, an experimental research design using a pre-test such as we described in the classical experimental design (shown in Figure 5-4) is considered to be a stronger design.

Repeated-Measurement Design

Naturally the pre-test comes before the experiment starts and the post-test comes afterward, but exactly how long before and how long afterward? Researchers seldom know for sure. Therefore, an experiment may contain several pre-treatment and post-treatment measures, especially when a researcher does not know how quickly the effect of the independent variable should be observed or when the most reliable pre-test measurement of the dependent variable should be taken. Figure 5-7 shows the layout for such an experiment.

An example of a repeated-measurement design would be an attempt to test the relationship between watching a presidential debate and support for the

FIGURE 5-7
Repeated-Measurement Design

	Pre-test				Post-test			
Experimental Group	R	M_1	M_2	$M_3 \ldots$	X	M_1	M_2	$M_3 \ldots$
Control Group	R	M_1	M_2	$M_3 \ldots$	X	M_1	M_2	$M_3 \ldots$

X = Experimental manipulation

R = Randomization

M = Measurements

candidates. Suppose we started out by conducting a classical experiment, randomly assigning some people to a group that watches a debate and others to a group that does not. On the pre- and post-tests we might obtain the following scores:

	Pre-debate Support for Candidate X	*Treatment*	*Post-debate Support for Candidate X*
Experimental Group	60	Yes	50
Control Group	55	No	50

These scores seem to indicate that the control group was slightly less supportive of Candidate X before the debate (that is, the random assignment did not work perfectly) and that the debate led to a decline in support for Candidate X of 5 percent (60 – 50) – (55 – 50).

Suppose, however, that we had the following additional measures:

	Pre-test				*Post-test*		
	First	*Second*	*Third*	*Treatment*	*First*	*Second*	*Third*
Experimental Group	80	70	60	Yes	50	40	30
Control Group	65	60	55	No	50	45	0

It appears now that support for Candidate X eroded throughout the whole period for both the experimental and control groups and that the rate of decline was consistently more rapid for the experimental group (that is, the two groups were not equivalent prior to the debate). Viewed from this perspective, it seems that the debate had no effect on the experimental group, since the rate of decline both before and after the debate was the same. Hence the existence of multiple measures of the dependent variable, both before and after the introduction of the independent variable, would lead in this case to a more accurate conclusion regarding the effects of the independent variable.

Multigroup Design

To this point we have discussed mainly research involving one experimental and one control group, although in a previous example an experiment included three experimental groups rather than a single experimental and single control group. In a **multigroup design** more than one experimental or control group is created so that different levels of the experimental variable can be compared. This is useful if the independent variable can assume several values or if the researcher wants to see the possible effects of manipulat-

FIGURE 5-8
Multigroup Experimental Design

	Pre-test		Post-test
R Experimental Group A	M_{exp1A}	X	M_{exp2A}
R Experimental Group B	M_{exp1B}	X	M_{exp2B}
R Experimental Group C	M_{exp1C}	X	M_{exp2C}
.	.	.	.
.	.	.	.
.	.	.	.
R Control Group	$M_{control1}$		$M_{control2}$

X = Experimental manipulation

R = Random assignment of subjects to groups

M = Measurements

ing the independent variable in several different ways. Multigroup designs may involve a post-test only or both a pre-test and a post-test. They may also include a time series component. In Figure 5-8 we show a diagram of the layout and analysis of this design. It uses the same notation as before, with R signifying randomization, *M* the various pre- and post-test measures, and so forth. Given the various experimental groups, one can make several comparisons among the levels of the independent variable.[11] (Read, for instance, M_{exp1A} as the "average measurement of members of experimental group A at the first time period, that is, pre-test.")

Here's an example. The proportion of respondents who return questionnaires in a mail survey is usually quite low. Investigators have attempted to increase response rates by including an incentive or a token of appreciation inside the survey. Since these items add to the cost of the survey, researchers want to know whether or not the incentives increase response rates and, if so, which ones are most effective and cost-efficient.

To test the effect of various incentives, we could use a multigroup post-test design. If we wanted to test the effects of five treatments, we could randomly assign subjects to six groups. One group would receive no reward (the control group), whereas the other groups would each receive a different reward—for example, 25¢, 50¢, $1.00, a pen, or a key ring. Response rates (the post-treatment measure of the dependent variable) for the groups could then be compared. In Table 5-3 we present a set of hypothetical results for such an experiment.

The response rates indicate that rewards increase response rates and that monetary incentives are more effective than are token gifts. Furthermore,

TABLE 5-3
Mail Survey Incentive Experiment

(Random Assignment)	Treatment	Response Rate (percent)
Experimental Group 1	25¢	45.0
Experimental Group 2	50¢	51.0
Experimental Group 3	$1.00	52.0
Experimental Group 4	pen	38.0
Experimental Group 5	key ring	37.0
Control Group	no reward	30.2

Note: Hypothetical data.

it seems that the dollar incentive is not cost effective, since it did not yield a sufficiently greater response rate than the 50¢ reward to warrant the additional expense. Other experiments of this type could be conducted to compare the effects of other aspects of mail questionnaires, such as the use of prepaid versus promised monetary rewards or the inclusion or exclusion of a pre-stamped return envelope.

Field Experiments

Laboratory experiments, whatever their power for making causal inferences, cannot be used to study a lot of (if not most of) the phenomena that interest political scientists. But some of the basic ideas of experimental design can be taken into the field.

Field experiments, or quasi-experiments, are experimental designs applied in a natural setting. As we noted, in a true experiment the investigator does two things: (1) randomly assigns participants to groups (for example, experimental and control), and (2) manipulates the experimental variable. In a field experiment, by contrast, there is no random assignment of participants to groups, but the investigator does try to manipulate one or more independent variables. The causal inferences are not as strong, but the design may be more practical.

As in any experimental research design, researchers attempt to control the selection of subjects and the manipulation of the independent variable. But in the field experiment the behaviors of interest are observed in a natural setting, increasing the likelihood that extraneous factors such as historical events will intrude and affect experimental results. Most important, because the subjects are not randomized, the groups do not necessarily start out the same in all relevant respects. Although it is possible to choose a natural setting that is isolated in some respects (and thereby approximates a controlled

environment), in general the researcher can only hope that the environment remains unchanged during the course of his or her experiment.

Still, field experiments should not be considered totally inferior to laboratory experiments. The artificial environment of a laboratory or controlled setting may seriously affect the external validity of a study's conclusions. Something that can be demonstrated in a laboratory may have limited applicability in the real world. Therefore, a program or treatment that is effective in a controlled setting may be ineffective in a natural setting. Field experiments are more likely to produce results that reflect the real-world impact of a program or treatment than are researchers' controlled experimental manipulations.

An interesting example of a field experiment in political science was the New Jersey experiment in income maintenance funded by the Office of Economic Opportunity and conducted from 1967 to 1971.[12] This effort was the forerunner of other large-scale social experiments designed to test the effects of new social programs. The experiment is a good illustration of the difficulty of testing the effects of public policies on a large scale in a natural setting.

At the time of the experiment, dissatisfaction with the existing welfare system was high because of its cost and because it was thought to discourage the poor from lifting themselves out of poverty. The system was also blamed for discouraging marriage and for breaking up families. Families headed by able-bodied men generally were excluded from welfare programs, and welfare recipients' earned income was taxed at such a rate that many thought there was little incentive for recipients to work.

In 1965 a negative income tax was proposed that provided a minimum, nontaxable allowance to all families and that attempted to maintain work incentives by allowing the poor to keep a significant fraction of their earnings. For example, a family of four might be guaranteed an income of $5,000 and be allowed to keep 50 percent of all its earnings up to a break-even point, where it could choose to remain in the program or opt out. If the break-even income was $10,000, a family earning $10,000 could receive a $5,000 guaranteed minimum plus half of $10,000 ($5,000) for a total of $10,000, or it could keep all the $10,000 earned and choose not to receive any income from the government. Critics of the proposed program argued that a guaranteed minimum income would encourage people to reduce their work effort. Others expressed concern about how families would use their cash allowances. Numerous questions about the administration of the program were raised as well. Because of these uncertainties, researchers designed the New Jersey income-maintenance experiment to test the consequences of a guaranteed minimum income system with actual recipients in a natural setting.

The research design included two experimental factors. One was the income guarantee level, expressed as a percentage of the poverty line. The level is the

TABLE 5-4
Experimental Design of New Jersey Income-Maintenance Experiment

Guarantee (percentage of poverty line)	Tax Rate		
	30%	50%	70%
125		X	
100		X	X
75	X	X	X
50	X	X	

Note: X represents experimental conditions actually tested.

amount of money a family received if other income was zero. The other factor was the rate at which each dollar of earned income was taxed.

Table 5-4 shows the experimental conditions of the two independent variables of interest to the researchers. Policy analysts were originally interested in income guarantee levels of 50, 75, 100, and 125 percent of the poverty level and tax rates of 30, 50, and 70 percent. The 4 × 3 **factorial design** displayed in Table 5-4 would allow researchers to examine the effect of the variation in one factor while the other was held constant and to measure the effects of different combinations of the independent variables. For example, it would allow researchers to examine the effect of varying the tax rate from 30 to 50 to 70 percent while the guarantee remained set at 75 percent of the poverty line.

Certain theoretical combinations of the experimental conditions were not chosen for study because they were unrealistic policy options or because they increased the cost of the study. Therefore, actual income-maintenance results were investigated for only eight of the twelve possible experimental conditions, and families participating in the study were assigned to one of those eight conditions. In assigning families to "cells" representing experimental treatments, there was a trade-off between the number of families that could be included in the study and the number of families assigned to each cell, since some cells were more costly than others. The cells representing the most likely national policy options (the 100–50 and 75–50 plans) were assigned more families to make sure that enough families completed the experiment. Finally, for some of the less generous treatments, the researchers experienced difficulty in finding eligible families willing to participate in the experiment. Families placed in these cells were likely to receive at most a small payment because they were near the break-even point. This situation created resentment within the community because all the families participating in the program had hoped to gain benefits beyond the nominal payment they could expect each time they completed an income report. If the researchers had had complete control over their subjects, assignment problems

would have been fewer. But in research involving human subjects, such control is understandably lacking.

Only families headed by able-bodied males were eligible for the experiment because of the great interest in the possible impact of the program on the work effort of poor families. Information about the work behavior of females with dependent children was not considered a good indicator of the work response of able-bodied males to public assistance. Very little was known about the work response of males because, as a group, able-bodied men and their families were generally not entitled to public assistance.

In the rest of this section we explore some of the issues and problems faced by the researchers during the course of this field experiment and discuss its outcome.

GENERALIZABILITY. To limit possible extraneous factors, families were originally chosen from a fairly homogeneous setting—New Jersey. Because a nationally dispersed sample was not chosen, however, the ability to generalize findings to a national program was limited. Generalizability was also affected by the three-year duration of the experiment. Families knew that the program was not permanent, and this may have affected their behavior.

INSTRUMENTATION DIFFICULTIES. The experiment also encountered problems with income measurement. Participants were asked to report their gross income, but families had trouble distinguishing between net and gross income. Families in the experimental groups learned how to fill out the reports correctly more quickly than did control group families because the experimental group families were asked to report income every month. Control group families (that is, other low-income families) were asked to report income only every three months. As a result, the accuracy of income data changed over time and differentially for experimental and control group families. This one-month/three-month difference arose because researchers were afraid that too much contact with control families would change their behavior (instrument reactivity) and make them less than true controls. This is an example of the trade-offs that researchers must make to avoid the numerous threats to the validity of experiments.

UNCONTROLLED ENVIRONMENT. In field experiments, unlike in laboratory experiments, researchers are not in complete control of subjects' environments. This point was illustrated dramatically during the New Jersey income-maintenance experiment. In the middle of the experiment, New Jersey adopted a public assistance program called Aid to Families with Dependent Children Unemployed Parent (AFDC-UP). Eligible families included those with dependent children and an unemployed parent, male or female. One reason that New Jersey had been chosen as an appropriate location for the income-maintenance experiment in the first place was precisely because it did not offer AFDC-UP. When it became available, however, AFDC-UP provided an attractive alternative to

some of the experimental cell conditions, and many families dropped out of the experiment.

Another problem arose because there were not enough eligible families in the New Jersey communities that were chosen to provide sufficient ethnic diversity. As a consequence, an urban area in northeastern Pennsylvania was included. However, the families in that area faced conditions that differed from those of the New Jersey families, and the families varied on some important characteristics, such as home ownership. One purpose of the study was to examine whether ethnic groups responded differently to the income-maintenance program. Because whites were represented mostly from one site, it became difficult to separate ethnic differences from site-induced differences.

Ethical Issues. Even though participation in the program was voluntary, the researchers were concerned about the effect of termination of the experiment on families that had been receiving payments. At the start of the experiment, families were given a card with the termination date of payments printed on it. Researchers debated tapering off payments and providing families with reminders as the end of the experiment approached. They decided to remind the families once toward the end, and research field offices remained open as referral agencies in case families needed help. But none requested help. Answers to a questionnaire three months after the last payment indicated that the experiment caused no serious adverse effects on the families that had participated.

Major Findings. Among white male heads of families receiving negative income tax payments, there was only a 5 to 6 percent reduction in average hours worked. For black male heads of families the average hours worked increased, although not significantly. For Spanish-speaking male heads of families the average hours worked decreased but also not significantly. Researchers were unable to explain this unexpected finding, and therefore it may be unreliable. Black working wives did not change their behavior, whereas Spanish-speaking and white working wives reduced their work effort considerably. Experimental families made larger investments in housing and durable goods than control families. There was also an indication that experimental families experienced increased educational attainment.

Because of the many difficulties encountered, the income-maintenance experiment failed to provide accurate cost estimates for alternative negative income tax plans or clear findings on the work disincentive of various tax rates. Because of these shortcomings, the experiment was not able to provide conclusive evidence in favor of or against a negative income tax plan.

The New Jersey income-maintenance experiment illustrates the difficulty of studying a significant political phenomenon both experimentally and in a natural setting. The researchers who conducted this experiment developed plausible, significant, and testable hypotheses and used an imaginative re-

search design to test those hypotheses. They identified the most interesting experimental treatments, attempted to assign their subjects to those treatments to accomplish pre-treatment equivalence, and conducted their experiment over a fairly lengthy period of time and in a natural setting to increase the external validity of their findings. Still, their efforts to isolate the effects of the independent variables in question were stymied by the real-world behavior of their subjects and by their inability to control completely both the experimental treatment and the environment in which it was operating. Researchers with fewer resources and even less control over both their subjects and the introduction of experimental treatments find it even more difficult to conduct meaningful experimental inquiries.

We have spent a considerable amount of time describing several experimental research designs to illustrate how experimental designs can help researchers draw appropriate conclusions about the effects of independent variables. Experimental designs are potentially useful because they allow researchers to isolate the effects of independent variables by controlling the assignment of subjects to experimental treatments, the introduction of the experimental stimulus itself, and the presence of extraneous influences. As a result, well-conducted experiments permit the evaluation of research hypotheses and the accumulation of causal knowledge.

Unfortunately, many of the sorts of hypotheses and behavioral phenomena of interest to political scientists do not lend themselves to the use of experimental research designs. Political scientists are limited by their inability to control completely the variables or the subjects of interest. Suppose, for example, that a researcher wanted to test the hypothesis that poverty causes people to commit robberies. Following the logic of experimental research, the researcher would have to randomly assign people to two groups, measure the number of robberies committed by members of the two groups prior to the experimental treatment, force the experimental group to become poor, and then at some later date measure again the number of robberies committed. Clearly, no researcher would be permitted to have this much control over a subject's life. Although the logic of experimental research designs is compelling, many researchers interested in explaining significant political phenomena have had to develop and employ other research designs.

Nonexperimental Designs

Because laboratory and field experiments are difficult to carry out, particularly when one wants to study aggregates like cities, counties, organizations, or countries, social scientists usually rely on nonexperimental approaches that are more practical. Although these methods are not as strong for making causal inferences, they allow the exploration of more realistic problems and

even the study of nonindividual units of analysis such as events, groups, and aggregates (for example, states or countries). Whatever the case, a **nonexperimental design** is a strategy for collecting information and data that will be used to test hypotheses and, if possible, make causal inferences. Such a design is characterized by at least one of the following: presence of a single group, lack of control over the assignment of subjects to groups, lack of control over the application of the independent variable, or inability to measure the dependent variable before and after exposure to the independent variable occurs. Because of these factors, causal inferences made using nonexperimental designs are not as strong as those possible through the classical randomized controlled experiment. Think of them as alternative plans or strategies for collecting data in a nonlaboratory setting.

We overview some possibilities in Table 5-5. Look, for example, at the row labeled "Surveys." A survey or poll can include anywhere from 100 to 5,000 (or more!) individuals, but polling fewer than 100 people is possible and not necessarily unsound. Furthermore, many research projects combine elements of different designs as in a panel study with an intervention interpretation (see below).

Whatever the design, the purpose of research is to collect information or data in order to test hypotheses, look for relationships among variables, and where possible make causal inferences. To compensate for the inferential shortcomings of the nonexperimental designs (especially the lack of random assignment), it is frequently necessary to achieve a rough approximation of randomization by statistical means. For instance, as demonstrated in later chapters, especially Chapters 12 and 13, surveys can be used to gather quantitative and qualitative data, which can then be manipulated mathematically to control for the effects of one or more extraneous variables while seeing how the main independent variable influences the dependent variable. A few of these designs are described in more detail here and in subsequent chapters. In reviewing them, we compare their features to the characteristics of ideal experiments mentioned earlier.

Small-N Designs

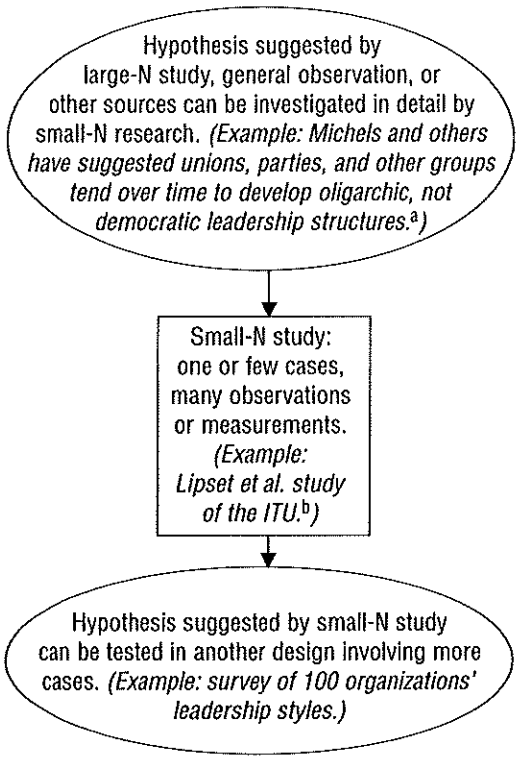
CASE STUDIES AND COMPARATIVE ANALYSIS. In a **small-N design** the researcher examines one or a few cases of a phenomenon in considerable detail, typically using several data collection methods, such as personal interviews, document analysis, and observation. When just one instance of a phenomenon is under investigation, the design is often called a *case study*; when two or more are involved the term *comparative* or *comparative case study* or analysis is frequently used. The units of analysis or the subjects of the study can be people

TABLE 5-5
Nonexperimental Research Designs

Design	Typical Number of Units or Cases (N)[a]	Examples of Units of Analysis	Purpose	Examples
Small-*N* Designs				
Single case (case study)	1	Event, nation, group, county, individual[b]	Usually one "thing" is studied in detail.	Study of French Revolution; study of U.S. House Ways and Means Committee
Comparative analysis	2 to 20	Events, nations, groups, counties, individuals	Two or more things are compared in relative detail.	Comparison of French and Russian revolutions; study of five Latin American countries; in-depth interviews with selected members of the British parliament
Focus group	10 to 20	Individuals	Often used in market research to probe reactions to stimuli such as commercials.	Test campaign ad's effectiveness
Cross-sectional Design				
Surveys (polls)	100 to 5,000	Individuals	A large number of people are measured on several variables to search for (possibly causal) relationships.	Study of U.S. public opinion and war in Iraq
Aggregate data analysis	20 to 500	Aggregates:[c] states, counties, cities, countries	Variables are often averages or percentages of geographical areas, but the goal is to search for (possibly causal) relationships.	Study of the death penalty and crime rates in different states; study of the relationship between union strength and welfare spending in developed countries
Longitudinal (Time-Series) Designs				
Trend analysis	20 to 300	Aggregates, individuals, cohorts[d]	Measurements on same variables at different time periods to examine changes in levels.	Study of changing levels of trust in government; level of unemployment and occurrence of civil strife in Europe, 1900–2000
Panel study	200 to 5,000	Individuals, households, cohorts	The *same* units are measured at different times to investigate relationships, changes in strength of relationships, and causality.	Study of changes in opinions toward prime minister of England
Intervention analysis	20 to 300	Individuals, cohorts, aggregates	Variables are measured at different times both before *and* after an intervention to see if it affected trends.	Study of number of traffic fatalities before and after enactment of seatbelt law in Texas

[a] These numbers are merely suggestive; some designs involve fewer or more cases.
[b] Biography, for example.
[c] Data are usually summations or averages of aggregations of individuals (often in geographical areas) such as median income in cities or counties.
[d] Individuals who experience the same event or experience or characteristics such as a birth cohort (people born in a specific year or period) or an event cohort (e.g., those who *first* voted in 1972).

FIGURE 5-9
Small-N Designs and Hypothesis Investigation



[a] Robert Michels, *Political Parties* (New York: Dover, 1959).
[b] Seymour Martin Lipset, Martin Trow, and James Coleman, *Union Democracy* (Garden City, N.Y.: Anchor, 1926).

(for example, prime ministers), events (such as the outbreak of the Korean War), institutions (for example, the U.S. Senate), nations or alliances (for example, NATO), decisions (such as the decision to invade Iraq), or policies (for example, *Roe v. Wade* dealing with abortion). The point is that one or a few cases or instances are studied in depth. As the sociologist Theda Skocpol explains, these types of designs involve "too many variables and not enough cases,"[13] meaning that the investigator collects lots of data on one or a few units.

A small-N study may be used for exploratory, descriptive, or explanatory purposes. Exploratory case studies are sometimes conducted when little is known about a phenomenon. Researchers initially may observe only one or a few cases of that phenomenon, and careful observation of this small set of cases may suggest possible general explanations for the observed behavior or attributes. These explanations—in the form of hypotheses—can then be tested more systematically by observing more cases. (See Figure 5-9.) Carefully scrutinizing the origins of political dissent within one group at one location may suggest general explanations for dissent; or following a handful of incumbent representatives when they return to their districts may suggest hypotheses relating incumbent attributes, district settings, and incumbent-constituency relations.[14]

Descriptive studies may be used to find out and describe what happened in a single or select few situations with a view toward finding avenues for further research. Here, the emphasis is not on developing general explanations for what happened. Alternatively, in some situations a single case may provide a critical test of a theory.[15] Recall from Chapter 2 that verification and falsification are crucial activities in science; finding a single exception may cast doubt on a previously accepted proposition. Therefore, if you can find a well-documented instance in which a widely accepted or important proposition does not hold, you may make a significant contribution. Finally, according to Robert K. Yin, case studies are most appropriately used to answer "how" or "why" questions.[16] These questions direct our attention toward explaining events. The strongest case studies start out with clearly identified theories that are expected to explain the events.

For years some scholars considered the use of case studies a suspect or even inferior research strategy, partly because of the limited sample sizes. Moreover, it might be thought that they are useless in causal analysis. But social scientists now recognize this type of design as a "distinctive form of empirical inquiry" and an important design to use for developing and evaluating public policies as well as for developing explanations for and testing theories of political phenomena.[17] Figure 5-9 illustrates some of these ideas. Suppose, for instance, a theory suggests that organizations such as labor unions tend to develop an oligarchic or elite leadership structure. But as we explain in a moment, a detailed investigation of a particular union can show if this proposition holds and if not, why not. By the same token, research on a single revolution might suggest propositions that could be verified later.

Proponents argue that a small-N design has some distinct advantages over experimental and cross-sectional designs for testing hypotheses under certain conditions. For example, a case study may be useful in assessing whether a statistical correlation between independent and dependent variables, discovered using a cross-sectional design with survey data (see below), is really causal.[18] By choosing a case in which the appropriate values of the independent and dependent variables are present, researchers can try to determine the timing of the introduction of the independent variable and how the independent variable caused the dependent variable. That is, they can learn whether a link exists between the variables and, therefore, can more likely offer an explanation for the statistical association. Benjamin Page and Robert Shapiro concluded their study of the statistical relationship between public opinion and public policy with numerous case studies.[19]

These studies differ from experimental designs in that the researcher is not able to assign subjects or cases to experimental and control groups nor can he or she manipulate the independent variable. Furthermore, the researcher does not control the context or environment as in a laboratory experiment. Yet the careful selection of a case or cases can lead to the approximation of a quasi-experimental situation. For example, a historian or political scientist may choose cases with different values of an independent variable but with the same values for important control variables. Cases with similar environments can be chosen. Lack of complete control over the environment or context of a phenomenon can be seen as useful. If it can be shown that a theory works and is applicable in a real situation, then the theory may more readily be accepted. This may be especially important, for example, in testing theories underlying public policies and public programs.

A good example of a case study is *Union Democracy* by Seymour Martin Lipset, Martin Trow, and James Coleman.[20] Sociologists and political scientists had long observed that voluntary organizations conform to Robert Michels's "iron law of oligarchy."[21] Lipset and his colleagues, however, observed that the

International Typographical Union (ITU) did not conform to the normal oligarchical pattern in which one group "controls the administration, usually retains power indefinitely, rarely faces organized opposition, and when faced with such opposition often resorts to undemocratic procedures to eliminate it."[22] The ITU had an institutionalized two-party system that regularly presented candidates for chief union posts elected in biennial elections. In *Union Democracy*, the authors attempted to explore this anomaly and in doing so helped explain the workings of democratic processes in general.

This kind of research may involve more than one case, in which instances they are often called *comparative case* studies. A comparative or multiple case study is more likely to have explanatory power than a single case study because it allows for replication; that is, it enables a researcher to test a single theory more than once. For some cases, similar results will be predicted; for others, different results will be predicted.[23] Multiple cases should not be thought of as a sample. Cases are not chosen using a statistical procedure to form a representative sample from which the frequency of a particular phenomenon will be calculated and inferences about a larger population drawn. Rather, cases are chosen for the presence or absence of factors that a political theory has indicated to be important.

As an example of the logic and layout of a comparative study and its potential utility in causal analysis, suppose a political scientist wanted to know why socialism never emerged as a major political force in the United States, especially in contrast to European nations such as Great Britain, a situation that has intrigued innumerable scholars for the past hundred years.[24] The most commonly cited "causes" of the failure of socialism to take root in America include "its relatively high levels of social equalitarianism, [enormous] economic productivity, and social mobility (particularly into elite strata), alongside the strength of religion, the weakness of the central state, the earlier timing of electoral democracy, ethnic and racial diversity, and . . . the absence of fixed social classes."[25] Imagine the researcher trying to sort out these possibilities by comparing France and the United States.

One strategy is the application of Mill's *method of difference* as introduced by the English philosopher John Stuart Mill: "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former: the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon."[26] For example, our investigator would first identify a country in which the condition (socialism) is present and one in which it is not. He or she would then look for similarities and differences in these antecedents. As the hypothetical data in Table 5-6 suggest, it might be that in the nineteenth century the United States and France shared

TABLE 5-6
Mill's Method of Difference

Case (Country)	Condition or Effect (Socialist Movements)	Antecedent 1 (Industrialized)	Antecedent 2 (Urbanized)	Antecedent 3 (Common Language)	Antecedent 4 (Historically Strong and Fixed Social Classes)
United States	yes	yes	yes	yes	no
France	no	yes	yes	yes	yes

similar experiences such as extensive industrialization and urbanization and that in both countries citizens spoke common languages (French and English) but they differed in that the French had a fixed and rigid social class system (including a landed aristocracy), whereas the United States did not. Since the two countries have parallel backgrounds except for their class structures, we may infer that this difference, rather than the other factors, explains why socialism has not had much of an influence on American politics.

Needless to say, this comparison is woefully inadequate and simplistic. In fact, no real analysis would take exactly this form. Instead, the table is a reconstruction (see Chapter 2) of the logic of comparison using this method. (Mill, by the way, introduced several other comparative methods, but knowing them is not essential for understanding the gist of comparative analysis.[27]) If you were to attempt research of this sort, you would have to consider many more factors and make difficult decisions about when an antecedent is or is not present. But the method of difference and similar designs underlies a great deal of political research.[28]

Despite the important contribution that case studies can make to our understanding of political phenomena, there are some concerns about the knowledge they generate.[29] One potential problem is the lack of rigor used in presenting evidence and the possibility for bias in using it. Typically, researchers sift through enormous quantities of detailed information about their cases. But how does one know all the important possible antecedents have been identified? Has something significant been omitted? Or the researcher may be the only one to record certain behavior or phenomena. Still, the potential for bias of this sort is not limited to case studies.

Another frequently raised criticism of case studies is the problem of generalization. One response to this criticism is to use multiple case studies. In fact, as Yin points out, the same criticism can be leveled against a single experiment—scientific knowledge is usually based on multiple experiments rather than on a single experiment.[30] Yet people do not say that performing a single experiment is not worthwhile. Furthermore, Yin states: "Case studies,

like experiments, are generalizable to theoretical propositions and not to populations or universes. In this sense, the case study, like the experiment, does not represent a 'sample,' and the investigator's goal is to expand and generalize theories (analytic generalization) and not to enumerate frequencies (statistical generalization)."[31]

A third potential drawback of case studies is that they may require long and arduous efforts to describe and report the results owing to the need to present adequate documentation. (Think about the complexity of untangling the differences between French and American societies.) This criticism may stem from confusing the case study with particular methods of data collection, such as participant observation (discussed in Chapter 8), which often requires a long period of data collection.[32] But case studies should not be ruled out as an appropriate research design due to this historic association.

Finally, in spite of the enthusiasm for case studies, considerable debate remains about just how strong causal inferences can be in these designs. Consider Table 5-6 again.[33] If our hypothetical study had discovered that France and the United States had the same values on *all* the independent variables, we would conclude that none of the hypotheses in this instance holds. And, as we stressed in Chapter 2, that might be an important conclusion given that falsification of propositions is one of the goals of science. But instead the conclusion seems to be that having a deep-seated social class system is at least a necessary condition for the emergence of socialism. Yet the result is hardly definitive. If, for example, the "real" cause of the dependent variable is not identified and explicitly included in the analysis, this design cannot detect it. Or it is plausible that the nonexplanatory variables (for example, industrialization, common language) interact or have a simultaneous joint effect on the dependent variable. In a design of the sort illustrated in Table 5-6 it is impossible to know.[34] Furthermore, this argument assumes causation is deterministic: once X appears, Y *always* follows. Yet, many, if not most scholars regard probabilistic causation—if X appears, then Y *probably* follows—as a more realistic description of the way the world works. Is it possible that deep social-economic cleavages do not always produce socialist movements? The method of difference provides no foolproof answer.[35]

Still, in many circumstances the case study design can be an informative and appropriate research design. The design permits a deeper understanding of causal processes, the explication of general explanatory theory, and the development of hypotheses regarding difficult-to-observe phenomena. Much of our understanding of politics and political processes comes from case studies of individual presidents, senators, representatives, mayors, judges, statutes, campaigns, treaties, policy initiatives, and wars. The case study design should be viewed as complementary to, rather than inconsistent with, other experimental and nonexperimental designs.

FOCUS GROUPS. A focus group consists of a small number of individuals (about twenty, say) who meet in a single location and discuss with a leader a topic or research stimulus such as a proposed campaign brochure. A focus group can superficially resemble an experiment, but usually no effort is made to assign participants randomly to treatment and control groups or to systematically introduce an experimental variation. The deliberations may or may not be (surreptitiously) recorded or observed by others on the research team. This approach lends itself nicely to market research but for the reasons just mentioned is seldom used to make causal inferences.

Focus groups have become somewhat controversial in politics because, critics assert, the results often help candidates, groups, and parties take safe or noncontroversial positions on issues. Some seemingly daring policy proposals, for example, have been thoroughly researched in focus groups, which eased the minds of political consultants that their candidates stood very little risk in adopting them. Yet, these small group discussions can be used to create hypotheses that can then be tested in larger surveys. Suppose you wanted to conduct a poll on your campus about physician-assisted suicide. You might begin with a focus group discussion to see generally what students thought about the issue. The verbal reports might then assist you in developing specific items for a questionnaire.

Cross-sectional Designs: Surveys and Aggregate Data Analysis

Perhaps the most common nonexperimental research design is cross-sectional analysis. In a **cross-sectional design,** measurements of the independent and dependent variables are taken at approximately the same time,[36] and the researcher does not control or manipulate the independent variable, the assignment of subjects to treatment or control groups, or the conditions under which the independent variable is experienced. If the units of analysis are individuals, the study is often called a *survey* or poll; if the subjects are geographical entities such as states or nations or other groupings of units, the term *aggregate analysis* is frequently applied. In either situation the units are simply measured or observed and the data recorded. The measurements are taken at roughly the same time (in an ideal world they would be taken simultaneously). In surveys the respondents themselves report their exposure to various factors. In aggregate analysis the investigator only observes which units have what values on the variables. There is no physical assignment of subjects to treatments. The measurements are used to construct, with the help of statistical methods, post-treatment quasi-experimental and quasi-control groups. Here we use the term *quasi* to indicate that the researcher does not control the assignment to the experimental or control groups. Rather the researcher sorts subjects into naturally occurring groups based on the subjects'

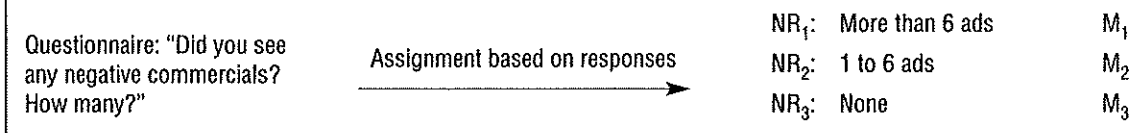
values on an independent variable, and the measurements of the dependent variable are used to assess the differences between these groups. Data analysis, rather than physical manipulation of variables, is the basis for making causal inferences.

Although this approach makes it far more difficult to measure the causal effects that can be attributed to the presence or introduction of independent variables (treatments), it allows observation of phenomena in more natural, realistic settings; increases the size and representativeness of the populations studied; and allows the testing of hypotheses that do not lend themselves easily to experimental treatment. In short, cross-sectional designs improve external validity at the expense of internal validity.

The example presented at the beginning of the chapter illustrates a particularly simple cross-sectional study. Recall that we tried to assess the effects of negative campaigning on the likelihood of voting by interviewing (that is, surveying) a sample of citizens and then categorizing respondents according to their answers to questions such as "Did you happen to see or read any campaign ads? How many? How many seemed to attack the opponent?" We could then sort respondents by their self-reported level of exposure to negativity. Because there is no random assignment (only self-reports), we do not control who is in each group by forcing people to view differing levels of negative advertising. The groups are simply observed. Figure 5-10 shows the layout of this study. If the groups differed by their rate of voting, we would have identified a relationship but would not demonstrate cause and effect.

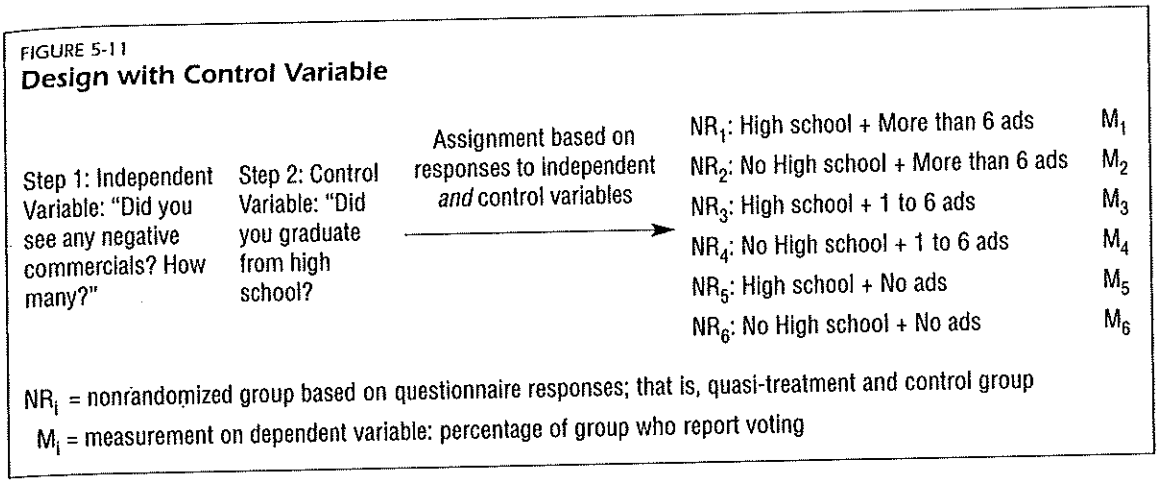
Suppose, for instance, we find that M_1 (the percentage of those who viewed six or more negative ads who also reported voting) is less than M_2, which in turn is less than M_3. Because of our research design and our inability to ensure that those with less and those with more exposure were alike in every other way, we could not necessarily conclude that campaign tone determines the propensity to vote. And this is true no matter how large our sample. With a survey design, then, we typically have to use data analysis techniques to control for potential confounders that may affect both the independent and dependent variables. If we wanted to control for these factors, we would have to in-

FIGURE 5-10
Logic of Survey Design



NR_i = nonrandomized group based on questionnaire responses; that is, quasi-treatment and control groups
M_i = measurement on dependent variable: percentage of group who report voting

FIGURE 5-11
Design with Control Variable



NR_i = nonrandomized group based on questionnaire responses; that is, quasi-treatment and control group

M_i = measurement on dependent variable: percentage of group who report voting

clude appropriate questions in the survey and then use statistics to hold them constant. Suppose we thought that education independently affected the propensities to watch a lot of television and to not vote. We would include in the survey a question about the level of the respondents' schooling, as indicated in Figure 5-11. Here we have formed six, nonrandomized groups and can compare their participation rates, M, as before. Presumably if education is creating the (spurious) relationship between voting and viewing habits, the M values for those with the same level of education would all be about the same except for sampling and measurement error. If, however, advertising does affect motivations even after controlling for education, the average measurements in the groups would vary. Under the hypothesis being considered, M_1 and M_2 would be less than, say, M_3 and M_4.

It might be possible to analyze this same problem with cross-sectional aggregate data, but we would confront the same difficulties. For example, we might treat "campaigns for the House of Representatives" as the unit of analysis. Then for each campaign in our survey we would try to determine (1) the proportion of radio, print, and television ads that attack the opposition—presumably we would have an objective standard for judging advertisements as negative, positive, or neutral—and (2) the percentage of the electorate, in each district, that cast a ballot. We might find that turnout was substantially lower when campaigns used a lot of negative advertising (say, more than 50 percent), compared with those races involving less negativity. Again, we would identify an association—the more negative the campaign, the lower the participation. But, and this is the essential point, we would be unable to infer causality. It is possible, after all, that the "negative" House races *also* or coincidentally took place in regions where turnout is generally low because of historical and economic factors, whereas the "clean" elections occurred where levels of participation are generally higher. Of course, we could control for past turnout, but

that would not eliminate the possibility that some other, unmeasured variable is creating a spurious relationship between campaign tone and voting.

In essence, the limitations of the cross-sectional design—that is, lack of control over exposure to the independent variable and inability to form pure experimental and control groups—force us to rely on data analysis techniques to isolate the impact of the independent variables of interest. This process requires researchers to make their comparison groups equivalent by holding relevant extraneous factors constant and then observing the relationship between independent and dependent variables, a procedure described more fully in Chapters 12 and 13). Yet holding these factors constant is problematic, since it is difficult to be sure that all relevant variables have been explicitly identified and measured. Even if a causal variable is not recognized and brought into the analysis, its effects are still operative.

A more detailed and real example of these ideas is Edward Tufte's use of cross-sectional aggregate data in a study of whether compulsory automobile safety inspection programs help reduce traffic fatalities.[37] Tufte's hypothesis was that "states with inspection programs have fewer automobile deaths than states without inspection programs." He measured the relevant variables at the same time, even though he used the average of auto fatality rates in three years as the dependent variable. The average number of traffic fatalities per 100,000 people for states with mandatory inspections was 26.1. For states without inspections, it was 31.9. Hence, he observed, in our notation, that $M_{\text{quasi-experimental}}$ was less than $M_{\text{quasi-control}}$.

Given Tufte's data, would we be safe in concluding that inspection programs caused the lower death rate? Possibly, but there are some problems with this conclusion. First, because the study lacks a pre-test of the dependent variable, we don't know whether a difference has always existed in the auto death rates of these two groups of states. Even before adopting inspection programs, some states may have had very low death rates for reasons that have nothing to do with car inspections. (We discuss alternative approaches to this static design in the next section.) Second, because Tufte did not control the assignment of the states to the two groups, he could not be sure that all relevant extraneous factors were distributed at random. Tufte did statistically control for some relevant differences among states, yet states with inspection programs may still have differed systematically from states without programs in some other way that was related to traffic fatalities. Hence we cannot be certain that any portion of the difference in fatality rates between the two groups of states can be attributed to the effect of an automobile inspection program.

That said, the difference in average death rates between the two groups of states may in fact understate the benefits of inspection programs, especially if many of the inspection programs were weak or poorly implemented. This pos-

sibility is plausible, since the treatment given in the states with inspection was not controlled by Tufte and could not be carefully observed. Clearly the lack of a pre-test and of control over the assignment of cases to the quasi-experimental and quasi-control groups creates difficulties for researchers who use the cross-sectional design.

Let's wrap up this section by recalling from Chapter 3 the inferential problems raised by aggregate data analysis. To wit, conclusions based on aggregates may or may not apply to individuals. For example, political scientists Donald Matthews and James Prothro used data from the 1950s on counties in the American south and found that the level of voting registration of African Americans was negatively correlated with median years of schooling among whites.[38] That is, the higher the level of education in a county, the lower the proportion of blacks who were registered. The data, however, pertain to counties. And, as the authors knew full well, it can be misleading, even fallacious, to extrapolate to the behavior of individuals from information gathered on aggregations. In Chapter 3 we called this type of inference an ecological fallacy. The point is that aggregate data analysis can be extremely helpful for the study of many social and political topics. But you must always remember the units of analysis to which conclusions apply. If you have measured, say, counties, your conclusion will, strictly speaking, apply to them and not to individuals or other kinds of units.

Large Longitudinal (Time Series) Designs

Longitudinal or time series designs are characterized by the availability of measures of variables at different points in time. As with the other nonexperimental designs, the researcher does not control the introduction of the independent variable(s) and must rely on data collected by others to measure the dependent variable. On the other hand, they have two distinct advantages: (1) Change in the level of variables or conditions can be measured and modeled, and (2) it is sometimes easier to decide time order or which comes first, X or Y. For political scientists interested in the dynamics of political phenomena, these are important considerations. Being static, cross-sectional studies do not lend themselves as well to these purposes.

Additional benefits of longitudinal studies include the fact that they can in principle estimate three kinds of effects: age, period (history), and cohort. Age effects can be considered a direct measure of (chronological) time, and its effects can be assessed like other variables. As in cross-sectional work, an investigator may be interested in the effect of age on political predispositions or ideology. (It is commonly asserted that as people age they become more politically conservative. Presumably something in the aging process—experiences, changes in hormones—affects people's perceptions and attitudes.) But in longitudinal analysis a **period effect** may be thought of as an indicator of

history during a time period, and this "history," not chronological age, is what matters. (During the late 1960s, for example, events such as Watergate and the Vietnam War adversely affected many citizens' trust in government, whether they were young or old. When that era passed, the effects on newer generations dissipated. So people who lived through those stormy times might have much different beliefs and opinions than younger people.)

Another way of interpreting time series effects is to consider cohorts.[39] A **cohort** is defined as a group of people who all experience a significant event at roughly the same time. A birth cohort, for instance, consists of those born in a given year or period; an event cohort includes those who shared a common experience such as their first entry into the labor force at a particular time. It is often hypothesized that one cohort will, because of its shared background, behave differently from individuals in a different cohort. To take one example, people born in the years immediately after World War II (the baby boomers) may have different political attitudes and affiliations than those who were born in the 1980s. Note that cohort and period effects are inescapably related because "cohort (year of birth) = period (duration of an event) – age (years since birth)."[40] There are, in short, a number of ways of understanding longitudinal research. The choice depends on the analyst's interests.

TREND ANALYSIS. As the term implies the analysis of a trend starts with measurements on a dependent variable of interest taken at different time periods (usually 20 or more) and attempts to determine whether and why the level of the variable is changing. A simple example shows changes in party identification—the feeling of being "close" to one of the parties—over the last 30 years. Figure 5-12 indicates that the percentages of citizens identifying with the Democratic and Republican Parties have declined significantly. If you are a party activist or journalist, this information is important. Still, it doesn't explain *why* partisanship seems to be slipping.

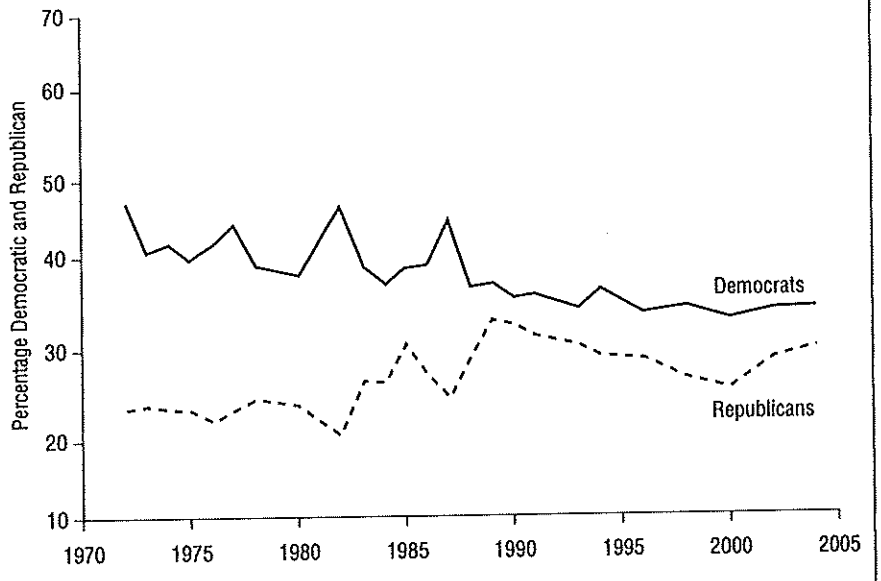
For that purpose an investigator needs to introduce additional variables and measure them over time. This type of analysis takes (roughly speaking) this form:

$$Y_1 = f(Y_{t-1} + Y_{t-2} \ldots + X^1s_t + X^1s_{t-1} + X^1s_{t-2} \ldots),$$

where the Y's and X's are measures of the dependent (Y) and independent (X) variables at the current (latest) time (t) and at previous times (t – 1, t – 2, . . .) and "*f*" means "is a function of . . ." or "is produced by . . ."[41] When data are measured at many different points, as illustrated in Figure 5-12, statistical procedures called time series analysis are often used. Although these techniques are somewhat advanced, they build on the ideas presented here and in later chapters.

FIGURE 5-12

Party Identification, 1972–2004



Source: General Social Survey 1972–2004 Cumulative Data File. Used with permission.
Note: Percentages of respondents who identify with a party strongly or moderately. Those who lean toward a party or are independent or apolitical have been excluded.

An example of this approach is Christopher Gelpi, Peter D. Feaver, and Jason Reifler's investigation of Americans' support for overseas military operations, described in Chapter 1. In particular, the authors noted that in April 2003 "76 percent of the public approved of President Bush's handling of the war; by the time of the U.S. election in November 2004, that number had dropped to 47 percent; a year later, it had dropped below 35 percent."[42] But instead of merely documenting this change, the authors attempted to explain it by examining trends in the numbers of American casualties and perceptions of progress in the war effort. Refer to Figure 1-2, on page 23. It graphs President Bush's approval ratings and military deaths in various phases of the war. Notice that the trend in the dependent variable (Bush's approval ratings) has been plotted along with trends in causality rates. Such a technique might have uncovered covariation among the variables. The data, however, did not back up the initial hypothesis that support would decline as casualties increased. This finding led to a more nuanced analysis in which the investigators examined the effects of changes in media coverage, perceptions of success, and events on the ground.

Note also that, although the previous examples pertain to changing proportions in samples of individuals, trends in aggregate variables (for example, crime or poverty rates in urban areas) can also be investigated.

Panel Studies

Suppose a public opinion analyst wants to learn if and how changes in a dependent variable, such as preferences for a particular candidate, are affected by changes in one or more independent variables, such as increasing attention to a campaign. A **panel study** is a cross-sectional design that introduces a time element. A researcher taking this approach measures the variables of interest on the same units of analysis at several different times. This strategy may be used to observe (cumulative or net) changes over time and to provide an analogue to a pre-test of some phenomenon prior to natural exposure to the experimental stimulus. A panel study is similar to a cross-sectional study, however, in that the subjects are measured at the same times in each "wave," and the researcher has no control over which subjects are exposed to the experimental stimulus.

Let us return to one of the hypothetical examples of a classical pre-test/post-test experiment described earlier in this chapter. In that example, we were interested in finding out whether or not exposure to a candidate's televised campaign commercials increased voters' ability to identify the important issues in a campaign. If we used the pre-test/post-test experimental design to test this hypothesis, we would measure pre-exposure issue awareness, randomly assign people to an experimental or a control group, expose only the experimental group to the commercials, and then measure post-exposure issue awareness again.

When using a panel research design, a researcher would proceed in a slightly different way. First, pre-exposure issue awareness would be measured for a group of subjects (presumably before any commercials have been broadcast). The researcher would wait for time to pass, the campaign to begin, and the commercials to be broadcast. Then the researcher would interview the same respondents again and measure both the amount of exposure to commercials and the post-exposure issue awareness for everyone. Finally, the researcher would use the measure of commercial exposure to construct quasi-experimental and quasi-control groups and compare the change in the amount of issue awareness for the two groups.

The major difference between the panel study and the classical experiment is that in the former the researcher waits for exposure to the experimental stimulus to occur naturally and then uses the amount of exposure reported by the respondents to create naturally occurring quasi-experimental and quasi-control groups. Hence, the researcher observes rather than controls exposure to the experimental stimulus.

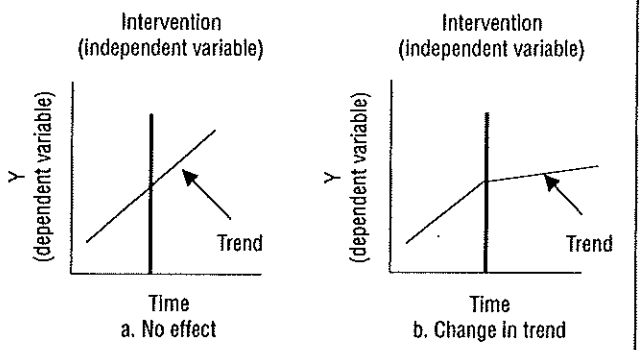
Because the panel study has a pre-test and a quasi-control group, the researcher can claim greater confidence in his or her conclusions than is possible with the cross-sectional design. However, the lack of control over who is exposed to the independent variable, and under what conditions, creates the problem of nonequivalent experimental and control groups. In our example, those who are naturally exposed to more commercials may be more interested in politics and hence more likely to develop issue awareness than those who are not, for reasons that have nothing to do with exposure to commercials.

Panel studies are particularly useful in studies of change in individuals over time. One difficulty with panel studies, however, is that individuals may die, move away, or decide to drop out of the study—what researchers refer to as **panel mortality.** If these persons differ from those who remain in the study, study findings may become biased and unrepresentative.

Panel studies have often been used in election campaigns to investigate the changes in voter beliefs, attitudes, and behavior that may be attributed to aspects of a campaign. A panel study of opinion change during the 1980 presidential campaign, for example, relied on surveys conducted with the same national sample of voting-age citizens in January/February, June, and September of 1980.[43] Larry M. Bartels was interested in the effect of media exposure during a presidential campaign on "each of 37 distinct perceptions and opinions regarding the presidential candidates, their character traits, their issue positions, the respondents' own issue preferences, and (in the case of incumbent Jimmy Carter) various aspects of job performance."[44] Since Bartels had available both January/February measures of the dependent variables, which were presumably unaffected by campaign news exposure, and later measures of the same variables after four and seven months of campaign coverage, change in voter perceptions and opinions could be analyzed. Measures of exposure to television network news and daily newspapers during the campaign allowed the creation of quasi-experimental and quasi-control comparison groups, and measures of other attributes, such as party identification, permitted statistical control of other significant political factors. As a result, the researcher was able to demonstrate significant media effects during a campaign with considerable confidence, even without experimental control over the introduction of the media exposure in question.

INTERVENTION ANALYSIS. In one version of a nonexperimental time series design, called **intervention analysis** or interrupted time series analysis, measurements of a dependent variable are taken both before and after the introduction of an independent variable. Here we speak figuratively: the occurrence of the independent variable is *observed,* not literally introduced or administered. We could observe, for instance, the annual poverty rate both before and after the ascension of a leftist party to see if regime change makes

FIGURE 5-13
Possible Effects of an Intervention



any difference on living standards. The pre-measurements allow a researcher to establish trends in the dependent variable that are presumably unaffected by the independent variable so that appropriate conclusions can be drawn about post-treatment measures. Refer to Figure 5-13. Panel a shows an increase in a dependent variable over time. (Suppose it is the poverty rate in metropolitan areas.) At a specific moment or period, an intervention takes place (perhaps the enactment of a job training program). But the trend line remains undisturbed: Y grows at the same rate before and after the "appearance" of the independent variable. In this case the intervention did not interrupt or alter the trend. (We would conclude, for example, that the program did not affect the increase in poverty.) Now consider panel b. It shows an increase in Y until the intervention occurs, at which point the growth in the trend begins to abate. In this instance the introduction of the factor appears to have caused the trend to flatten (for example, the introduction of the job training program slowed the growth in poverty).

Intervention analysis works best when the independent variable (that is, the intervention) occurs at a particular moment or during a fairly brief period, affects a dependent variable that is routinely measured, or is known about in advance so that appropriate pre-test measurements can be made. This design would work well if, as in the previous example, we wished to evaluate the impact of a new program or policy initiative. To take another case we might try to evaluate the impact of sobriety checks on alcohol-related traffic accidents in states by examining the number of such accidents in the years before and after the introduction of the checks. If we observed a decline in accidents, we might conclude that sobriety checks had been effective. But whether the checks caused the decrease would remain unclear because other unmeasured things (the age distribution of the population, perhaps) may also be changing during the time period under study. If so, we cannot know if the checkpoints, the other factor(s), or both are affecting the fatality rate. The problem, of course, lies in the fact that there is no control group with which to compare the unit or units of analysis that experienced the independent variable. The results of a time series can often be improved if the researcher can identify quasi-experimental and quasi-control groups and produce a time series of measurements of the dependent variable for each. In this way the researcher can have more confidence

that the observed shift in the dependent variable is the result of the introduction or presence of the independent variable.

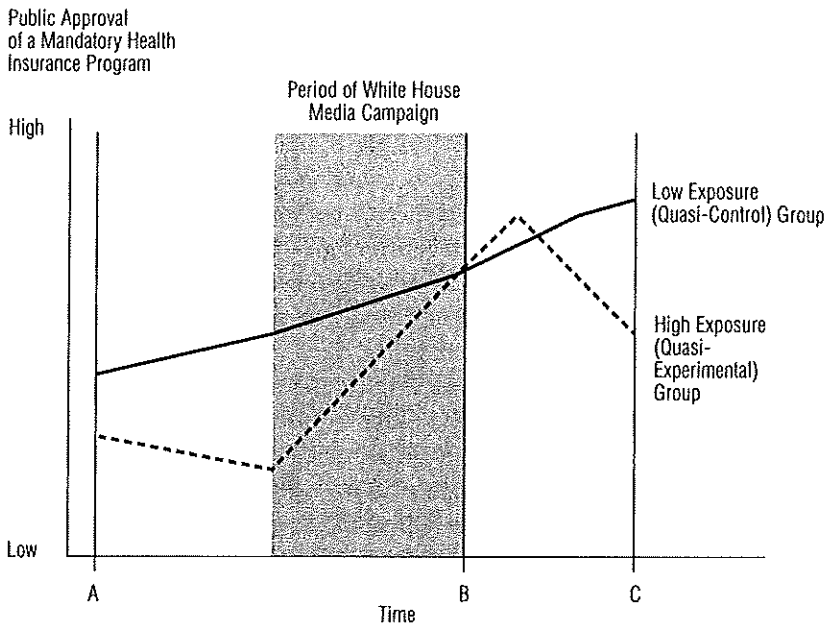
For example, some states may adopt health insurance programs for children whereas others do not. Researchers can compare children's health trends in both types of states using regularly collected indicators of children's health to determine whether or not the health of children in the states with insurance programs has improved relative to the health of those in states without the programs. Even though the researcher controls neither the assignment of states to the groups with or without the program nor the content or implementation of the programs, this situation is often referred to as a "natural" experiment because of the presence of before and after measurements for both quasi-experimental and quasi-control groups.

As another example, suppose that we are interested in whether or not an aggressive media campaign organized by an interest group has an effect on popular support for a public policy initiative, such as mandatory, comprehensive health care coverage. We might first obtain a series of public opinion polls measuring popular support for mandatory coverage. Using a measure of overall media exposure, we could then separate the respondents into two groups: those most likely to be exposed to the media campaign and those least likely to be exposed to the media campaign. By continuing the time series of popular support for health care during and after the media campaign and comparing the entire series for the quasi-experimental and quasi-control groups, we could assess the influence of the media campaign on popular support for the comprehensive health care initiative.

In Figure 5-14 we see the hypothetical results of such a time-series-with-quasi-control-group design. Before the introduction of the independent variable, the less exposed (quasi-control) group was more supportive of mandatory health care coverage, with a positive trend among the less exposed group and a negative trend among the more exposed group. (This could be because the less exposed group was more Democratic and less affluent than the more exposed group and because both groups were already responding to Washington and interest group rhetoric before the campaign began.) During the media campaign the downward trend in support in the quasi-experimental group was reversed, and by the end of the campaign the most exposed group was just as supportive of mandatory health care coverage as the less exposed group. After the media campaign concluded, the level of support among the most exposed group began to decline again, while the level of support in the less exposed group remained fairly constant. This is strong evidence that the media campaign had an effect, albeit one that diminished with time.

As with trend analysis, these types of problems are usually tackled with mathematical procedures, not graphs alone. Yet the pictures illustrate the logic underlying these more advanced methods.

FIGURE 5-14
Hypothetical Results of a Time-Series-with-Quasi-Control-Group Research Design: Effect of a White House Media Campaign



Alternative Research Strategies

To study a phenomenon such as the effects of the media on voters or the behavior of federal justices, political scientists most commonly use one of the experimental or nonexperimental designs described above. This propensity stems from their commitment to verify hypotheses empirically and strive for valid causal inferences. But these approaches hardly exhaust the list of possibilities. We now briefly describe two alternatives that flow from the quest for scientific knowledge but that rely on different tactics.

Formal Modeling

Anyone who has ever seen or perhaps built a model airplane knows full well that these replicas do not fly passengers or carry cargo or drop bombs. They are simply representations of reality. Nevertheless, they can be quite useful and not just for entertainment. At the very least they suggest what a real airplane looks like, and many can even be used for scientific purposes. Aeronau-

tical engineers, for example, use models to see how certain wing shapes affect a plane's stability or what a sudden downdraft will do to the wing's structural integrity. So, even though model planes may be totally unrealistic in one sense, they can still be useful, even essential, devices for learning about flight. Models, it turns out, are also quite useful in political science.

A **formal model** (frequently termed an *analytic model,* or just a *model*) is a simplified and abstract representation of reality that can be expressed verbally, mathematically, or in some other symbolic system, and that purports to show how variables or parts of a system tie together.† This definition may seem a bit vague, and so we describe the components of a model and then provide an example.[45]

The main parts of a model are (1) a set of "primitives," or undefined terms or words whose meaning is taken for granted, (2) a collection of assumptions, that is, statements or assertions whose validity is again taken for granted but are explicitly stated, (3) a body of rules or logic for linking the parts of the model together and making deductions, and (4) various derived propositions that are true by virtue of the rules used to deduce them. They may not be supported empirically, as we will see. A modeler has to start some place, so the most elemental parts of the system have to be taken as given. It is also necessary to make some assumptions. As an example, many political scientists assume—they do not prove with data—that people are "rational" in a sense to be discussed below. These primitives and assumptions are akin to the axioms used to construct geometries. What can make a model powerful is the application of rules, such as the rules of algebra or symbolic logic, that allow one to move from base terms and assumptions to validly derived statements. As noted, the conclusions in a model are true if the rules have been applied correctly, period. These statements may or may not hold in the real world. But most modelers believe that they will be approximately empirically valid and that, if the model is a good one, the understanding they provide more than compensates for inaccuracies in the predictions.

For an example of formal modeling let us return to the question of why people do or do not vote. This time we approach the question from a theoretical perspective. All models by definition simplify reality, and the model that follows is an especially unsophisticated or bare-bones version presented only for illustrative purposes.[46] Consider a single citizen. We assume that this person has desires or wants, which in modeling are often called "utilities." This individual may desire lower taxes, an end to federal gun control, and less military spending.‡ Denote the sum of the utilities as U. See? The model is already becoming abstract, and we are taking the meaning of "utility" more or less for granted. In any event, we can introduce some additional notation to clarify the amount of utility this person would get from two parties or candidates. For instance, let U_A be the value that Party (or Candidate) A brings to the voter on these three

issues *if* it comes to power. Similarly let U_B stand for the utility the other party (or candidate) brings if it takes office. Now the model *assumes* that voters are rational, which means that they try to maximize or get the most value or utility from their actions. This conception or definition of rationality is usually termed subjective or psychological because what is getting maximized are internal mental states such as expected happiness and contentment, states that are hard to measure objectively. It is as though people mentally calculate the difference between the utilities from A and B and use the result to guide their decisions: vote for A if U_A is greater than U_B, vote for B if U_B is higher, and abstain if U_A equals U_B. (Why? See below for the answer.) No one does exactly this, of course, but people may behave as though they did.[47] More formally, we might symbolize the comparison of utilities, which we can call a party differential, as follows:

$$U_{AVB} = U_A - U_B$$

Note this important point: if U_{AVB} equals zero, the person sees no difference between the parties or candidates (the utilities from each are identical), has no incentive to vote, and thus abstains.[48]

To this point we have introduced a few primitive terms (for example, *utility*) and one key assumption—namely, voters are rational, meaning that they vote for the alternative that brings them the most pleasure, or in the modeler's language, they maximize utility. From these elements we use the rules of algebra to derive predictions of how a person will behave. (It is just simple algebra that if U_A equals U_B, their difference is zero. And, by assumption, if an action brings zero utility, it is not taken.) True, the conclusion appears rather trivial—people vote for their most preferred party—but we can expand the model to reach a startling conclusion.

The model implicitly assumes that people have preferences and can easily act on them. But that presumption may be too simplistic because it does not take into account so-called information and transaction costs. That is, potential voters have to take time to find out where the parties stand on the relevant issues. That task may seem relatively trivial, but it's not. Politics has to compete with a lot of matters of importance to voters, including job and family matters, desire for relaxation and entertainment, and health issues. To become informed about electoral politics requires at the least a fair amount of time spent reading or listening to information about the campaign. There may

THE MEANING OF *VALIDITY*

It is necessary to separate standards of evaluation for formal models and experiments. With experiments we try to conduct research in such a fashion that we can make valid causal inferences based on empirical observation. The type of validity involved in models is different. In this case we say deductions are valid if and only if they have been properly deduced from the premises. These deductions may or may not be empirically valid. Just as the claim that the sum of the angles in a right triangle is 180 degrees follows from the axioms in Euclidean geometry and not the state of nature, derivations of a modeling process depend on starting premises and logic and not empirical reality. As we will see, we may want to test a model's predictions. If we find that they do not hold, it does not necessarily mean the model is incorrect, just that it does not describe the real world.

even be monetary costs, as in the expense of newspapers and magazines. In addition, parties and candidates frequently obscure their stands on issues or try to distort those of their opponents. And to make matters worse, potential voters have to inconvenience themselves to get registered, find the polling place, and take the time to actually vote.[49] A lot of these costs may not seem like much, but research tells us that they probably affect political behavior. So let us factor in a term C for the cost of becoming informed and going to the poll. Admittedly, this is an abstract term, since people do not literally summarize their expenses in a single number. Still, it does represent symbolically what people must *feel* when allocating their time and energy to certain tasks.

Now we can make the prediction that if C is greater than or equal to the absolute value of U_{AVB}, our hypothetical citizen will feel that the cost of voting outweighs any utility derived from one candidate's bringing more utility than another. In symbols, if

$$|U_{\text{AvB}}| \leq C,$$

then the costs of voting exceed the benefits that one or the other party (or candidate) brings and again we predict abstention. The rationality assumption (that is, that people maximize utility) explains why: if the costs of acting outweigh the benefits, action is not rational.

And there is another possible difficulty that we can model. Everyone must know that in any sizable election the probability that any one vote will be decisive is minuscule. For example, how likely is it that any one person's vote will determine the outcome of a contest for state representative that brings thousands of people to the polls? Elections are just not that close. (How many elections are decided by a margin of 5,001 to 5,000?) So anyone must reason that it is not going to make a huge difference in the outcome if he or she stays home on election day. That is to say, the chances of reaping the benefits from a favorite party's taking power are not affected by anything any particular voter does. And our citizen's participating or not will in all likelihood not affect a favorite party's chances of winning or losing. There are just too many votes being cast for a single vote to be decisive.

To see what taking this small probability into account does to the model of turnout, let the likelihood that a vote is decisive be P. This number will be exceedingly small in any realistic election, say one in ten thousand. And if we discount a person's utility derived from one party over another, the result will also be exceedingly small, almost approaching zero, as a matter of fact. The discounting (or multiplication) of utility by P (that is, $P|U_{\text{AvB}}|$) is called the expected or subjectively expected utility of A versus B.

Example: If the expected benefit of A over B is 100 units of utility, and the probability of casting the deciding vote and thus actually bringing about that

amount of utility is one in 10,000, then the result is 0.01. Now this should be compared with the cost of voting. After all, doesn't everyone compare expected gains with the costs of obtaining them? Any cost of becoming informed and voting over that amount, say 1 unit of utility, will mean the rational voter has no incentive to participate. And since we assume all voters are rational, and that all make similar calculations, we deduce that *no one* will ever vote!

Once again this result may seem far-fetched. But it is a conclusion of many models of voting and has even earned the title "the paradox of voting."[50] This paradox has been so troubling that it has occupied dozens and dozens of social scientists who have tried to figure out why this logically deduced (from the primitives and assumptions of the model) flies in the face of the reality that many, many people do, in fact, vote.

Perhaps we can rescue our model by introducing another kind of utility that goes above and beyond that obtained by seeing one party or another elected. This additional value might come from the pleasure one receives just by participating in politics and knowing that widespread apathy could undermine democracy.[51] Let us call this new factor E, which means the extra utility or value brought by being active in civic life. It enters the potential voter's "calculation" as an additional or extra utility. Table 5-7 summarizes our model.

We can interpret the table a couple of ways. The deductions follow logically from the premises. They are true despite what happens in the real world. If you are troubled by this situation, you might maintain that the model has a kind of internal validity, but its external validity is low because the results do not generalize to any meaningful population. This is a reasonable argument, but it brings us to a discussion of the value of formal models in political science.[52]

Many social scientists think that modeling as a research method has many advantages. Models, they believe, lead to clear and precise thinking. As Morris P. Fiorina puts it, modeling requires that we put "all the cards on the table."[53] For a model to be useful, definitions have to be unambiguous and assumptions made explicit. If these conditions hold and the rules are known and accepted, other researchers can verify the deductions. We may not like the way a term has been defined or an assumption introduced. But if we at least know what the model says, we can suggest alternatives and find and correct errors. Verbal theories or histories, by contrast, often contain hidden or vaguely defined terms and assumptions, and because of the ambiguities, people often talk past one another. Models also tend to be compact. They do away with all but the essential aspects of a problem. True, they may oversimplify, but simplification may be necessary when studying complex political phenomena. Besides, almost any kind of research involves such narrowing of problems to manageable proportions and the reliance on simplifying postulates, a point sometimes forgotten by those critical of this technique.

TABLE 5-7
Simple Formal Model of Turnout Based on Rationality for Citizen X

Primitive terms

- U_{AvB}: utility or value Citizen X gets from voting for Party (Candidate) A versus Party (Candidate) B.
- *P:* Citizen X's estimated probability that his or her vote will be decisive.
- *C:* the cost to Citizen X of becoming informed, registered, and voting.
- *E:* benefits to Citizen X of democracy and participating in elections.

Assumption

- Rationality: citizens act to maximize utility.

Predictions

- Decide to vote or not:
  1. $P|U_{AvB}| + E > C$ — Vote (see 3–4 below)
  2. $P|U_{AvB}| + E \leq C$ — Do not vote, but go to 6–7 below

- Decide how to vote (if 1 above holds):
  3. $U_{AvB} > 0$ — Vote for A
  4. $U_{AvB} \leq 0$ — Vote for B
  5. $U_{AvB} = 0$ — Do not vote, but go to 6–7 below

- If parties do not differ ($U_{AvB} = 0$):
  6. $E > C$ — Toss coin in voting booth (because $|U_{AvB}| = 0$)
  7. $E \leq C$ — Do not vote

Finally, political scientists apply formal modeling techniques in a surprisingly wide variety of areas, from international relations (for example, coalition formation, the outbreak of war, arms races) to the behavior of organizations and groups (legislative and judicial decision making, roll call voting, and committee assignments in Congress) to individual behavior (candidate preferences, candidate campaign tactics). Indeed, the procedures now occupy a prominent place in many professional journals and graduate school curricula.

We conclude this section by noting that despite the growing popularity of modeling in the social sciences, it has numerous critics who are especially upset at the use of assumptions to make a model "work." The rationality premise, for instance, causes concern because it is defined in ways that many find odd. Is a chain smoker who buys cigarettes at the lowest possible cost rational? Formal modelers would say "yes, if he or she acts to maximize (subjective or psychological) utility." But probably no doctor would agree.[54]

Simulation

A research design closely connected to modeling is simulation. For our purposes we define a **simulation** as a representation of a system by a device in

order to study the system's behavior over time. The units of analysis are not discrete individuals like those in a sample survey. Rather, the fundamental interest lies in a process or structure, such as a large organization, a legislative body, or a party system, which has several or many components. To the extent that individuals enter into a simulation, their behavior as a collectivity (for example, a crowd, a committee, a coalition) is the main interest. The device for investigating the phenomenon can be a computer program, a board game, role playing, or some other method that allows the investigator to see how the system's components interact and change. At least for investigative purposes the system or process is usually thought of as "closed," or not subject to external forces.

A major consideration in simulation studies is time. They emphasize the dynamic interaction of the internal elements. If one part changes, what happens to the others? Are there feedback loops or paths of reciprocal causation? Does the system evolve to an equilibrium state, or does it spin out of control? Consequently, simulations are normally "run." From a starting point or set of "givens" the pieces are made to interact in order to see two things: how each affects and is affected by the others, and what happens to the overall state of the system at different time intervals, or iterations.

Creating simulations requires knowing as much about the subject matter as possible. They are, therefore, not very helpful for exploratory work but are useful for discovering emergent properties that cannot be found in static models or easily calculated by conventional means. If the investigator has a good idea of a system's constituents and how they interrelate but cannot easily predict what happens when they start functioning together over a period of time, a series of runs may provide insights not available by computation or deduction. As a result, they work best or are most informative when a researcher has a solid, extensive body of knowledge with which to work.

You may be familiar with simulations, although not by that name. Some games, like Monopoly, are simulations. Role-playing exercises such as mock parliaments or model courtrooms are simulations that teach participants legislative or judicial behavior. But simulations are also widely used for academic and applied purposes to study complex phenomena that cannot be investigated with laboratory or survey tools. Many simulations appear in the press, as in models of the economy constructed by economists in an attempt to determine future levels of employment or inflation. In making these models they set certain parameters and then let one or more factors vary. What will happen, they might ask, if people suddenly decide to retire early? What will happen to productivity? To health care costs? To Social Security trust funds? As another example, the Congressional Budget Office, a nonpartisan research arm of Congress, routinely tries to predict the likely effects of alternative policy options (a tax cut, perhaps) on the state of government spending

or on the economy as a whole. The natural sciences, as you may know, frequently resort to simulations to investigate complex phenomena such as the effect of accumulating greenhouse gases on global climate.

To appreciate the possibilities, let's take a simplified example based on research conducted by Richard J. Stoll.[55] For generations, international relations theorists have wondered if and how in a world of self-interested, independent nations (a "Hobbesian world," it is sometimes called), a balance of power could be "automatically" achieved and maintained, especially in the absence of an external authority like a world government. For this study we assume the world consists of several dozen nations of varying power, all of which are locked in a struggle for survival. We want to know what will happen in this kind of an anarchic system if states start attacking one another. Will they eventually reach equilibrium or a balance of power? Will they mutually annihilate each other? Will one country come to dominate the others? These are interesting questions, because, after all, everyone lives in a world of sovereign states having different amounts of power and interests, and the potential for deadly conflict is always present.

The mechanism we use to assemble and "operate" the simulation depends on our purposes, the needed calculations, and similar considerations. For this example, we will use a computer program. This approach allows us to try different configurations and initial values, insert random errors, and run hundreds or even thousands of trials to see what on average happens. We cannot, of course, make an actual simulation here, but in Figure 5-15 we sketch a flowchart of the sequential steps our hypothetical program would go through on each iteration. A flowchart does not literally describe the inner workings of the machine, but it does suggest how a simulation operates.

We start by "creating" a set of countries and assigning them initial values. In this instance we might, as Stoll did, let the computer set up a table with, say, five rows and ten columns to make fifty internal countries. The program would also give each nation, which is just a storage location in the computer, values for different variables, such as its beginning power. (At the outset, the power scores might be allocated randomly.) We would also write computer programming code to reflect the rules of the game. Only contiguous states, for example, can instigate attacks on one another. These program statements represent what we think we know about an anarchic international system. If we have been woefully unrealistic, critics can tell us and adjustments can be made. The model described here takes type of regime for granted, but it might be desirable to write a rule to prevent democracies from attacking one another.[56]

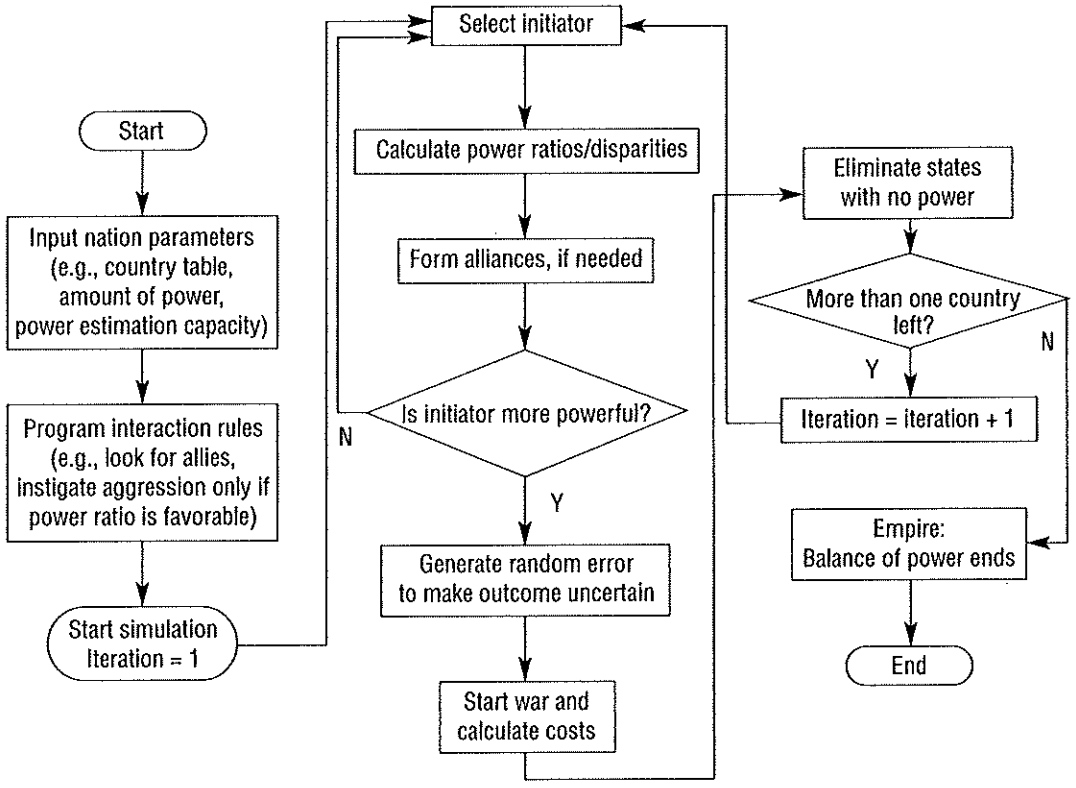
Then the simulation begins with the first round or iteration. An aggressor is selected at random, its power in relation to its neighbor is calculated, alliances are formed, and power is recalculated. Based on the rules a decision

is made to instigate war. The damage (cost) to the participants is calculated next, and new power scores allotted according to the rules. Some countries may be eliminated in the process. Now the program makes a decision: is more than one nation left? If yes, the simulation goes to the next round, during which an initiator is selected and new calculations are made. We might let the maximum number of iterations be any number, such as 500 or 1,000. Should countries remain after this maximum number of trials, we would conclude that the system has been preserved. If, however, only one state remains, the game ends with total domination by one superpower. The process can be run again and again to see what happens on average. Stoll based his study on 270 replications.

It is probably apparent that a simulation is something like a "thought experiment" in which a researcher wonders "what would happen if A does ___

FIGURE 5-15

Balance-of-Power Simulation



Source: Adapted from Richard J. Stoll, "System and State in International Politics," *International Studies Quarterly* 31 (December 1987): 387–402. Reprinted with permission of Blackwell Publishing.

and then B responds with ____." Such an undertaking may seem sterile or fruitless, but that is only because we have provided just a bare sketch of what is and can be done with simulations. As we noted, they have become an integral part of the social sciences.[57] A great deal of what we think we know about the world comes partly from simulation studies.

Conclusion

In this chapter we discussed why choosing a research design is an important step in the research process. A design enables the researcher to achieve his or her research objectives and can lead to valid, informative conclusions. We presented two basic types of research designs—experimental and nonexperimental—along with a couple of alternative approaches. We discussed their advantages and disadvantages. Experimental designs—which allow the researcher to have control over the independent variable, the units of analysis, and their environment—are often preferred over nonexperimental designs because the former enable the researcher to establish sounder causal explanations. Therefore, experimental designs are generally stronger in internal validity than are nonexperimental ones. However, it may not always be possible or appropriate to use an experimental design. Thus nonexperimental observation may also be used to test hypotheses in a meaningful fashion and often in a way that increases the external validity of the results. In these instances, causal assertions rest on weaker grounds and frequently have to be approximated by statistical means (see Chapters 12 and 13). Yet the basic objectives of research designs, whether experimental or nonexperimental, are the same.

A single research design may not be able to eliminate all threats to internal and external validity. Researchers often use several designs together so that the weaknesses of one can be overcome by the strengths of another. Also, findings based on research with a weak design are likely to be accepted more readily if they corroborate findings from previous research that used different designs.

Notes

*. If you have trouble following this idea, imagine that you have a large can of marbles, most of which are red but a few of which are blue. Now, draw randomly from the can a single marble and put it in a box. Then draw another marble—again randomly—and put this one in a second box. Repeat this process nineteen more times. At the end you should have two boxes of twenty marbles each. If you have selected them randomly, there should be approximately the same proportion of red and blue marbles in *each* box. If you started with a can holding 90 percent red marbles and 10 percent blue, for example, each of the two boxes should hold about eighteen red marbles and two blue ones. These may not be the exact numbers, but the boxes would differ only slightly—one might have three blue marbles and the other just one—but these differences will be due solely to chance.

†. The word *model* used alone can be misleading because social scientists employ it for many different purposes. But in this context we give it a particular meaning, namely, the activities described in the discussion that follows.

‡. In this version of rationality, the wishes do not have to be logically consistent. The desire for lower taxes may be inconsistent with less military spending. But political scientists generally take preferences as given. This inconsistency is a sore point with some critics of formal models.

1. CBS News, "Going Negative," September 27, 2004. Retrieved January 17, 2004, from www .cbsnews.com/stories/2004/09/27/politics/main645881.shtml.
2. See, for example, Stephen Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino, "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88 (December 1994): 829–838.
3. CBS News, "Going Negative." Also see Paul Freedman and Ken Goldstein, "Measuring Media Exposure and the Negative Effects of Campaign Ads," *American Journal of Political Science* 43 (October 1999): 1189–1208.
4. See Chapter 13 for a more thorough discussion of spurious relationships.
5. Ansolabehere, Iyengar, Simon, and Valentino, "Does Attack Advertising Demobilize the Electorate?"
6. See Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand-McNally, 1966), 5–6; and Paul E. Spector, *Research Designs* (Beverly Hills, Calif.: Sage Publications, 1981), 24–27. Four components of an ideal experiment are identified by Kenneth D. Bailey in *Methods of Social Research* (New York: Free Press, 1978), 191.
7. A good reference is Donald T. Campbell and David A. Kenny, *A Primer on Regression Artifacts* (New York: Guilford Press, 1999), especially chap. 3.
8. Martin T. Orne, "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications," *American Psychologist* 17 (November 1962): 776–783.
9. This discussion is based on Bailey, *Methods of Social Research,* 204–206; and Jarol B. Mannheim and Richard C. Rich, *Empirical Political Science* (Englewood Cliffs, N.J.: Prentice-Hall, 1981), 76–77.
10. Shanto Iyengar, Mark D. Peters, and Donald R. Kinder, "Experimental Demonstrations of the 'Not-So-Minimal' Consequences of Television News Programs," *American Political Science Review* 76 (December 1982): 848–858.
11. A major subfield in applied statistics is the design and analysis of this and more complicated types of experiments. For some introductions, see Steven R. Brown and Lawrence E. Melamed, *Experimental Design and Analysis* (Thousand Oaks, Calif.: Sage Publications, 1998); and Larry Toothacker, *Multiple Comparisons* (Thousand Oaks, Calif.: Sage Publications, 1990). One of the best books is the classic by Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand-McNally, 1966).
12. This discussion is based on Joseph A. Pechman and P. Michael Timpare, eds., *Work Incentives and Income Guarantees: The New Jersey Negative Income Tax Experiment* (Washington, D.C.: Brookings Institution, 1975), esp. chaps. 2 and 3.
13. Theda Skocpol, *States and Social Revolutions* (New York: Cambridge University Press, 1979), 36.
14. See Richard F. Fenno Jr., *Home Style: House Members in Their Districts* (Boston: Little, Brown, 1978).
15. Robert K. Yin, *Case Study Research: Design and Methods,* rev. ed. (Beverly Hills, Calif.: Sage Publications, 1989), 47.
16. Ibid., 17–19.
17. Ibid., 21.
18. Alexander L. George, "Case Studies and Theory Development: The Method of Structured, Focused Comparison," in Paul Gordon Lauren, ed., *Diplomacy: New Approaches in History, Theory and Policy* (New York: Free Press, 1979), 46; and Skocpol, *States and Social Revolutions,* chap. 1.
19. Benjamin Page and Robert Shapiro, "Effects of Public Opinion on Policy," *American Political Science Review* 77 (March 1983): 186.

20. Seymour Martin Lipset, Martin Trow, and James Coleman, *Union Democracy* (Garden City, N.Y.: Anchor, 1962).
21. Robert Michels, *Political Parties* (New York: Dover, 1959).
22. Lipset, Trow, and Coleman, *Union Democracy*, 1.
23. Yin, *Case Study Research*, 53.
24. See, among countless other sources, Werner Sombart, *Why There Is No Socialism in the United States* (White Plains, N.Y.: International Arts & Sciences Press, 1976; first published in German in 1906); and Seymour Martin Lipset, *The First New Nation: The United States in Historical and Comparative Perspective* (New York: Basic Books, 1963).
25. Seymour Martin Lipset and Gary Marks, *It Didn't Happen Here: Why Socialism Failed in the United States* (New York: Norton, 2000): 16.
26. John Stuart Mill, *Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation* (New York: Harper & Brothers Publishers, 1858) 225. (Available on Google Book Search, http://books.google.com/, accessed February 11, 2007.)
27. See Merrilee H. Salmon, *Logic and Critical Thinking*, 2d ed. (San Diego: Harcourt Brace Jovanovich, 1989), 109–115.
28. Skocpol's *States and Social Revolutions* is an excellent example of seminal research that explicitly uses Mill's method.
29. Yin, *Case Study Research*, 21–22.
30. Ibid., 21.
31. Ibid., 23.
32. Ibid.
33. The literature discussing the pros and cons of small-N research, especially its applicability to causal inference includes, among many others, Gary King, Robert Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton: Princeton University Press, 1991); James Mahoney, "Strategies of Causal Analysis in Small-N Analysis," *Sociological Methods and Research* 28 (May 2000): 387–424; and Stanley Lieberson, "Small N's and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases" *Social Forces* 70 (December 1991): 307–320.
34. Lieberson, "Small N's and Big Conclusions," 312–313.
35. For an extended discussion, see King, Keohane, and Verba, *Designing Social Inquiry.* For a more optimistic picture of the possibilities of causal inference in small-N research, see Douglas Dion, "Evidence and Inference in Comparative Case Study," *Comparative Politics* 30 (January 1998): 127–145.
36. Although the measurements may be taken over a period of days or even weeks, cross-sectional analysis treats them as though they were obtained simultaneously.
37. Edward R. Tufte, *Data Analysis for Politics and Policy* (Englewood Cliffs, N.J.: Prentice-Hall, 1974), 5–17.
38. Donald Matthews and James Prothro, "Socio-Economic Factors and Negro Voter Registration in the South," *American Political Science Review* 57 (March 1963): 36–38.
39. The relationships among age, period, and cohort can be expressed as follows: Cohort (e.g., year of birth) = period (e.g., duration of an event) – age (e.g., years since birth). See Scott Menard, *Longitudinal Research* (Newbury Park, Calif.: Sage University Paper Series on Quantitative Applications in the Social Sciences, 1991), 7.
40. Ibid.
41. In practice, relationships of this sort are thought of as probabilistic, not deterministic, so a random error term would be added.
42. Christopher Gelpi, Peter D. Feaver, and Jason Reifler, *International Security* 30 (Winter 2005/06): 8–9.
43. Larry M. Bartels, "Messages Received: The Political Impact of Media Exposure," *American Political Science Review* 87 (June 1993): 267–285.

44. Ibid., 269.
45. This discussion is based on Morris P. Fiorina, "Formal Models in Political Science," *American Journal of Political Science* 19 (February 1975): 133–159. Also see Michael Laver, *Private Desires, Political Action: An Invitation to the Politics of Rational Choice* (Beverly Hills, Calif.: Sage Publications, 1997).
46. A classic work that builds a more thorough model of voter turnout (and many other political phenomena) is Anthony Downs, *An Economic Theory of Democracy* (New York: Harper and Row, 1957). Also see William H. Riker and Peter C. Ordeshook, "A Theory of the Calculus of Voting," *American Political Science Review* 62 (March 1968): 25–42.
47. Isn't this just a formalization of statements such as "Well, on the whole I just prefer Bush over his opponent?"
48. This deduction comports with the commonly heard complaint from nonvoters: "There ain't a dime's worth of difference between ____ and ____."
49. Not everyone in society can bear these costs equally. College graduates may find them less onerous than high school dropouts, which explains in part why the less educated vote less regularly than do those with more education.
50. John A. Ferejohn and Morris P. Fiorina, "The Paradox of Not Voting: A Decision Theoretic Analysis," *American Political Science Review* 68 (June 1974): 525–536.
51. In *An Economic Theory of Democracy,* Downs added exactly this sort of ad hoc term into his model. He called it the "long-run participation value."
52. Donald P. Green and Ian Shapiro provide a major critique of formal models in *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science* (New Haven: Yale University Press, 1994). Their analysis particularly faults many models for leading to trivial conclusions and empirically invalid predictions. These points have been rebutted by several social scientists in Jeffrey Friedman, ed., *The Rational Choice Controversy* (New Haven: Yale University Press, 1996).
53. Fiorina, "Formal Models in Political Science," 137.
54. Green and Shapiro, *Pathologies of Rational Choice Theory,* and Jeffrey Friedman, *The Rational Choice Controversy,* explore this issue in greater detail.
55. Richard J. Stoll, "System and State in International Politics: A Computer Simulation of Balancing in an Anarchic World," *International Studies Quarterly* 31 (December 1987): 387–402.
56. Although we are presenting this simulation for illustrative purposes and do not care much about its realism, the last sentence reminds us of an important point. Science is a cumulative process that builds and rebuilds on the work of others. If we have stated our model's components clearly, then others can see where it might be weak, and we or they might make additions and corrections. In this instance, international relations theorists have noted that democracies generally do not go to war with one another, and these theorists have devoted much time and effort trying to figure out why not. Whatever the answer, we might want to explicitly include this consideration in our simulated international system.
57. For discussions and examples of social science simulations, see Robert Axelrod, "Advancing the Art of Simulation," in Rosario Conte, Rainer Hegelsmann, and Pietro Terna, eds., *Simulating Social Phenomena* (New York: Springer, 1997): 21–40; and Nigel Gilbert, "Simulation: An Emergent Perspective," Centre for Research in Social Simulation, n.d. (based on lectures given in 1995 and 1996), www.soc.surrey.ac.uk/research/cress/resources/emergent.html.

Terms Introduced

ASSIGNMENT AT RANDOM. Random assignment of subjects to experimental and control groups.

CAUSAL RELATIONSHIP. A connection between two entities that occurs because one produces, or brings about, the other with complete or great regularity.

CLASSICAL EXPERIMENTAL DESIGN. An experiment with the random assignment of subjects to experimental and control groups with a pre-test and post-test for both groups.

COHORT. A group of people who all experience a significant event in roughly the same time frame.

CONTROL GROUP. A group of subjects who do not receive the experimental treatment or test stimulus.

CROSS-SECTIONAL DESIGN. A research design in which measurements of independent and dependent variables are taken at the same time; naturally occurring differences in the independent variable are used to create quasi-experimental and quasi-control groups; extraneous factors are controlled for by statistical means.

DEMAND CHARACTERISTICS. Aspects of the research situation that cause participants to guess the purpose or rationale of the study and adjust their behavior or opinions accordingly.

EXPERIMENTAL EFFECT. Effect, usually measured numerically, of the independent variable on the dependent variable.

EXPERIMENTAL GROUP. A group of subjects who receive the experimental treatment or test stimulus.

EXPERIMENTAL MORTALITY. A differential loss of subjects from experimental and control groups that affects the equivalency of groups; threat to internal validity.

EXPERIMENTATION. A research design in which the researcher controls exposure to the test factor or independent variable, the assignment of subjects to groups, and the measurement of responses.

EXTERNAL VALIDITY. The ability to generalize from one set of research findings to other situations.

EXTRANEOUS FACTORS. Factors besides the independent variable that may cause change in the dependent variable.

FACTORIAL DESIGN. Experimental design used to measure the effect of two or more independent variables singly and in combination.

FIELD EXPERIMENTS. Experimental designs applied in a natural setting.

FORMAL MODEL. A simplified and abstract representation of reality that can be expressed verbally, mathematically, or in some other symbolic system, and that purports to show how variables or parts of a system are interconnected.

HISTORY. A change in the dependent variable due to changes in the environment over time; threat to internal validity.

INSTRUMENT DECAY. A change in the measurement device used to measure the dependent variable, producing change in measurements; threat to internal validity.

INSTRUMENT REACTIVITY. Reaction of subjects to a pre-test.

Internal validity. The ability to show that manipulation or variation of the independent variable actually causes the dependent variable to change.

Intervention analysis. Also known as interrupted time series design. A nonexperimental design in which the impact of a naturally occurring event (intervention) on a dependent variable is measured over time.

Maturation. A change in subjects over time that affects the dependent variable; threat to internal validity.

Multigroup design. Experimental design with more than one control and experimental group.

Nonexperimental design. A research design characterized by at least one of the following: presence of a single group, lack of researcher control over the assignment of subjects to control and experimental groups, lack of researcher control over application of the independent variable, or inability of the researcher to measure the dependent variable before and after exposure to the independent variable occurs.

Panel mortality. Loss of participants from a panel study.

Panel study. A cross-sectional study in which measurements of variables are taken on the same units of analysis at multiple points in time.

Period effect. An indicator or measure of history effects on a dependent variable during a specified time period.

Post-test. Measurement of the dependent variable after manipulation of the independent variable.

Precision matching. Matching of pairs of subjects with one of the pair assigned to the experimental group and the other to the control group.

Pre-test. Measurement of the dependent variable prior to the administration of the experimental treatment or manipulation of the independent variable.

Randomization. The random assignment of subjects to experimental and control groups.

Repeated measurement design. An experimental design in which the dependent variable is measured at multiple times after the treatment is administered.

Research design. A plan specifying how the researcher intends to fulfill the goals of the study; a logical plan for testing hypotheses.

Selection bias. Bias in the assignment of subjects to experimental and control groups; threat to internal validity.

Simple post-test design. Weak type of experimental design with control and experimental groups but no pre-test.

Simulation. A simple representation of a system by a device in order to study its behavior.

Small-N design. Type of experimental design in which one or a few cases of a phenomenon are examined in considerable detail, typically using several data collection methods, such as personal interviews, document analysis, and observation.

SPURIOUS RELATIONSHIP. A relationship between two variables caused entirely by the impact of a third variable.

STATISTICAL REGRESSION. Change in the dependent variable due to the temporary nature of extreme values; threat to internal validity.

SURVEY RESEARCH. The direct or indirect solicitation of information from individuals by asking them questions, having them fill out forms, or using other means.

TEST STIMULUS OR FACTOR. The independent variable introduced and controlled by an investigator in order to assess its effects on a response or dependent variable.

TESTING. Effect of a pre-test on the dependent variable; threat to internal validity.

Suggested Readings

Campbell, Donald T., and Julian C. Stanley. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand-McNally, 1966.

Cook, Thomas D., Donald T. Campbell, and Thomas H. Cook. *Quasi-Experimentation: Design and Analysis Issues.* New York: Houghton Mifflin, 1979.

Creswell, John W. *Research Design: Qualitative and Quantitative Approaches.* Thousand Oaks, Calif.: Sage Publications, 1994.

Downs, Anthony. *An Economic Theory of Democracy.* New York: Harper and Row, 1957.

Hakim, Catherine. *Research Design: Strategies and Choices in the Design of Social Research.* London: Allen and Unwin, 1987.

Hanley, Ryan Patrick. "What Is a Case Study and What Is It Good For?" *American Political Science Review* 98 (May 2004): 327–354.

Laver, Michael. *Private Desires, Political Action: An Invitation to the Politics of Rational Choice.* Thousand Oaks, Calif.: Sage Publications, 1997.

Menard, Scott. *Longitudinal Research.* Newbury Park, Calif.: Sage University Paper Series on Quantitative Applications in the Social Sciences, 1991.

Sambanis, Nicholas. "Using Case Studies to Expand Economic Models of Civil War," *Perspectives on Politics* 2 (June 2004): 259–279.

Sekhon, Jasjeet S. "Quality Meets Quantity: Case Studies, Conditional Probability, and Counterfactuals" *Perspective on Politics* 2 (June 2004): 281–293.

Spector, Paul E. *Research Designs.* Beverly Hills, Calif.: Sage Publications, 1981.

Vendaele, Walter. *Applied Time Series and Box-Jenkins Models.* New York: Academic Press, 1983.

Yin, Robert K. *Case Study Research: Design and Methods,* rev. ed. Beverly Hills, Calif.: Sage Publications, 1989.