CHAPTER 3

Research Design

What are the effects of negative political advertisements? 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? 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 the appropriate "units of analysis" (e.g., people, nations, states, organizations) are appropriate for the tests; what measurements or data are needed; how all this information will be collected; and the best analytical and statistical procedures. All the parts of a research design should work to the same end: drawing sound conclusions that are supported by evidence.

David Nachmias and Chava Nachmias define a research design as a plan that "guides the investigator in the process of collecting, analyzing, and interpreting observations. It is a model of proof that allows the researcher to draw inferences concerning causal relations among the variables under investigation... Furthermore, [it] also defines the domain of generalizability; that is, whether the obtained interpretations can be generalized to a larger population or to different situations."

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 things affect the choice of a particular design. One is the purpose of the investigation. Whether the research is intended to be exploratory, descriptive, or explanatory will most likely influence the choice. 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. Consequently, many common designs that researchers actually use entail unfortunate but necessary compromises, and consequently 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 by researchers 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 briefly 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. Consequently, 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, in a campaign flooded with negative ads turnout will be lower 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 aware of attack ads on television, 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 those who were not aware of these ads all 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. (We will see the reason for using these letters in a moment.) What this table symbolizes is a relationship or association between X and Y.

This strategy is frequently called opinion 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 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 3.1.) 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, it is possible that those who missed the ads differ in other ways as well from those who saw them. Perhaps they have a higher level of education and that accounts for their higher turnout rate. Or maybe they had a generally strong sense of civic duty and 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 what we call a "spurious," or false, 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 3.1 illustrates causal and spurious relationships.2

Distinguishing real, causal relations 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 inferences. Ideally such a design does three things.

- Covariation. First, it demonstrates 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 3-1, viewing negative advertisements is connected to not voting, 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.

- Time order. Second, 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 note that however reasonable this assumption we have not really demonstrated it empirically. And 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.

- Elimination of possible alternative causes, sometimes termed “confounding factors.” Finally, 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 3-2 shows the possibilities presented by the third requirement. The first diagram (Causal Relationship) shows a “true” causal connection between 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 there is an apparent association that 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, while those with less are not as apt to go to the polls. (The arrow with the positive sign [+] means positive causal effect of education 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." 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 later in Chapter 10.) This approach is perhaps the most common in the social sciences, and it is the one followed in the hypothetical example above. 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 Experiment**

Experimentation allows the researcher to control exposure to an experimental variable (often called a test factor or 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.3 (A survey, as we discuss below and in Chapter 10 can be considered another kind of research design. But we also use survey "questionnaires" within experiments to collect data on opinions and beliefs.) This situation results partly from the nature of the phenomena of greatest interest.
to political science, such as who votes in actual elections. Nevertheless, it is important to understand experimental design because it provides a standard of 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 the following five basic characteristics of a classical randomized experimental design:

- First, the experimenter himself or herself establishes two groups: an experimental group (actually, there can be more than one) that receives or is exposed to an experimental treatment, or test stimulus or factor, and a second called the control group because its subjects do not undergo the experimental manipulation. So, for example, Anselsbehere 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 leaving others alone.

- 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 males and females, 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.

- Third, 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.

- Fourth, 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.

Finally, 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 in order to see if negative political advertising depresses the intention to vote. This case is purely hypothetical, but it resembles the research conducted by Anscombe and his associates, and more to the point it shows the inferential power of experiments. (The example will also show 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 afterward we find a difference in propensity to vote, how will we be able to tell if it arose because of the advertisement or gender? We illustrate the procedure in Figure 3-3.

![Figure 3-3: A Randomized Controlled Experiment]

Note: R = randomization.
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 randomly pick them. 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 him 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 (e.g., age, sex, family income, years of education, place of birth) and about their political beliefs and opinions (e.g., 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 35.5 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 9, “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. To disguise our purpose 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 an identical fifteen-minute news broadcast. So far both groups have been treated the same. If there are any differences between them, they are the result of happenstance.

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 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 3-2.

Note, first, that 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 3-2 we show that the two factors 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 alternative explanation of the covariation, for our randomization and experimental manipulation ensured that the groups differed (on average) 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.

In order to simplify later explanations we introduce a general framework for describing research designs, as shown in Figure 3-4. We let $X$ stand for an experimental manipulation (e.g., 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 just 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 should be about zero.
FIGURE 3.4

**Layout of Randomized Controlled Experiment**

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R</strong> Experimental Group</td>
<td>$M_{exp1}$</td>
<td>$X$</td>
</tr>
<tr>
<td><strong>R</strong> Control Group</td>
<td>$M_{control1}$</td>
<td>$M_{control2}$</td>
</tr>
</tbody>
</table>

- $X$ = Experimental manipulation
- $M$ = Measurements
- Experimental effect = $(M_{exp2} - M_{exp1}) - (M_{control2} - M_{control1})$
- $R$ = Random assignment of subjects to groups

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 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 actually 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 "cancel 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.
**Research Design**

With regard to those factors that are known to influence the outcome or suspected of doing so. An especially important measure in this regard is $M_{\text{exp}} - M_{\text{control}}$, which should be about zero.

And, 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 there are many factors 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. In setting up experimental and control groups, social scientists generally lack sufficient authority or incentives to offer subjects.

**Interpreting and Generalizing the Results of an Experiment**

Most readers, we hope, have followed the logic of our arguments. But they must be flabbergasted at the unreality of the hypothetical example we introduced and wonder how anyone could make a definitive statement about negative advertising based on these data, even if we had actually carried out this experiment on real people using real television newscasts and commercials. Someone might exclaim, "This test is invalid." It may be, but before jumping to that conclusion we need to consider carefully and closely the term "validity."

Statistical theory tells us that experiments properly conducted can lead to valid conclusions about causality. In this context, however, "validity" has a particular meaning, namely, that the manipulation of the experimental or independent variable itself, and not some other variable, did in fact influence the dependent variable. Social scientists call this kind of validity, internal validity. **Internal validity** means the research procedure demonstrated a true
cause-and-effect relationship that was not created by spurious factors. Social scientists generally believe that the type of research design we have been discussing—a randomized controlled experiment—has strong internal validity. But it is not foolproof.

Several things can affect internal validity. As we have argued, the principle strength of experimental research is that the researcher has enough control over the environment to make sure that exposure to the experimental stimulus is the only difference between experimental and control groups and that at the outset all comparison groups have the same traits, except for sampling error. Sometimes, however, history, or events other than the experimental stimulus that occur between the pre-test and post-test measurements of the dependent variable, will affect the dependent variable. For example, suppose that after being selected and assigned a room the experimental subjects happen to hear a radio program that undercuts their faith in the electoral process. Such a possibility might arise if there was a long lag between the first measurement of their attitudes and the start of the experiment. This situation is shown schematically in Figure 3-5. In this instance no one would be able to say if the radio program (Z in the figure) or the experimental treatment produced the effect on voting intentionality.

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

---

**FIGURE 3-5**

**The Effects of "History" on an Experiment**

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>H Experimental Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R Control Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X = Experimental manipulation</td>
<td>$M_{\text{exp1}}$</td>
<td>$Z$</td>
</tr>
<tr>
<td>R = Randomization</td>
<td>$M_{\text{control}}$</td>
<td>$Z$</td>
</tr>
<tr>
<td>Z = Nonexperimental event</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M$ = Measurements</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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. Consequently, 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 biases can also lead to problems. Such biases 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 may 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.

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 mortality, 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. There may be something about our experiment on political advertising that tips off subjects that we, the researchers, expect to find that negative ads depress turnout, and so they (perhaps unconsciously) adjust their feelings in order to prove the proposition and hence please us. In this case it is the desire to satisfy the researchers' objectives that affects the disposition to vote, not the commercials themselves. 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 disguise to both patients and attendants who is receiving a real experimental medicine and who is receiving a traditional medicine or placebo.

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 that can 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 3-3, 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, the 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.
RESEARCH DESIGN

Threats to the external validity of an experiment also may be caused 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 "real" voters are exposed to dozens and dozens of spot 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, when subjects know they are being studied they may react to a stimulus differently from 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. Before considering alternative approaches let us consider another actual 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. To test the effect of the hypothesized independent variable (exposure to television news), the researchers employed 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 participants 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) actual 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 external validity issues discussed above. 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. Consequently, 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.

Note that one might argue that the second part of their experiment was not actually 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 pretest and post-test does not absolutely establish the factors that caused the change. The researcher can never be sure what might have happened to subjects if no treatment had been given at all. Nevertheless, there may be good reasons for omitting a control group. Because the researchers had already conducted an experiment in which there was 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 will briefly describe some variations on this approach. Each one represents a different attempt to retain experimental control over the experimental situation while also dealing with threats to internal and external validity. Although you may not have an opportunity to employ these designs, knowledge of them will help in understanding published research and in determining whether the research design being employed supports the author's conclusions.

Simple Post-test Design

The simplest 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 3-6.

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 simple example. Suppose we wanted to test the hypothesis that watching a national nominating convention on television makes 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 experimental or 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 pretreatment differences, an experimental research design using a pre-test such as we described in the classical experimental design (shown in Figure 3.4) is considered to be a stronger design.

**Time Series 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 experimental times series design, a research design that includes several pre-treatment and post-treatment measures, may be used when a researcher is uncertain exactly 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.

An example of an experimental time series design would be an attempt to test the relationship between watching a presidential debate and support for the 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 watch the debate. On the pre- and post-tests we might receive the following scores:

<table>
<thead>
<tr>
<th></th>
<th>Pre-debate Support for Candidate X</th>
<th>Post-debate Support for Candidate X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental Group</td>
<td>60</td>
<td>Yes</td>
</tr>
<tr>
<td>Control Group</td>
<td>55</td>
<td>No</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
<td>Second</td>
</tr>
<tr>
<td>Experimental Group</td>
<td>80</td>
<td>70</td>
</tr>
<tr>
<td>Control Group</td>
<td>65</td>
<td>60</td>
</tr>
</tbody>
</table>

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
Research Design

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 groups are 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 manipulating 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 3-7 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.9

Here's an example. The proportion of respondents who return questionnaires in a mail survey is usually quite low. Consequently, investigators have attempted to increase response rates by including an incentive or token of

<table>
<thead>
<tr>
<th>FIGURE 3.7</th>
<th>Multigroup Experimental Design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-test</td>
</tr>
<tr>
<td>R Experimental Group A</td>
<td>$M_{exp1A}$</td>
</tr>
<tr>
<td>R Experimental Group B</td>
<td>$M_{exp1B}$</td>
</tr>
<tr>
<td>R Experimental Group C</td>
<td>$M_{exp1C}$</td>
</tr>
<tr>
<td>R Control Group</td>
<td>$M_{control1}$</td>
</tr>
<tr>
<td>X = Experimental manipulation</td>
<td></td>
</tr>
<tr>
<td>R = Random assignment of subjects to groups</td>
<td></td>
</tr>
<tr>
<td>M = Measurements</td>
<td></td>
</tr>
</tbody>
</table>
appreciation inside the survey. Since these add to the cost of the survey, researchers want to know whether or not the incentives really do 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 3-3 we present a set of hypothetical results for such an experiment.

The response rates indicate that rewards increase response rates and that monetary incentives have more effect than do token gifts. Furthermore, 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

As might be readily guessed, 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
Research Design

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, the subjects are not randomized and so 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. 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 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 1969 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 experimental design included two experimental factors. One was the income guarantee level, expressed as a percentage of the poverty line. The level is the 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 3-4 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 x 3 factorial design displayed in Table 3-4 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 more quickly how to fill out the reports correctly 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 was dramatically illustrated 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 thus 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 different conditions from those of the New Jersey families and 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,
Research Design

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 discussed above, 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 is a good example of 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 employed an imaginative research 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 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 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.

**Causal Inference in 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 have developed some nonexperimental approaches that are more practical. Most of these approaches—but not all—involve soliciting information from subjects or respondents via a questionnaire. But whatever the case, a nonexperimental 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. Although nonexperimental designs can be applied to a wide variety of topics, they do not lead to as strong causal inferences as experiments do. Think of them as alternative plans or strategies for collecting data in a nonlaboratory setting.

- Survey research, as we saw earlier, means gathering information about the characteristics, behavior, or attitudes of a relatively large group of people, often referred to as a "population." The main tool is the administration of a questionnaire to subjects who respond by choosing among specific alternatives or by giving open-ended verbal answers. When done for commercial purposes this approach is often called "market research"; newspapers usually refer to it as a (public opinion) poll. Keep in mind that survey methodology is frequently thought of as a "kind" of research design, as when an investigator decides to use a poll instead of an experiment to investigate a problem. But the measurements used in experiments also frequently rely on questionnaires to solicit information. We discuss surveys in greater detail in Chapter 10.

- 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 no effort is usually made to assign participants randomly to treatment and control groups. The deliberations may or may not be (surreptitiously) recorded or observed by others on the research team.
RESEARCH DESIGN

- Almost as common, political scientists collect “aggregate” data that describe collectivities (e.g., precincts, states, countries) from various sources like census reports, national archives, or previous studies. The study of determinants of “winning and losing” in politics that we mentioned provides an example of this type of analysis.

- Investigators sometimes turn to content or document analysis to collect information. For example, follow-up studies to the Anslobehrere investigation of negative advertising discussed earlier analyzed the tone and text of many actual campaign commercials in order to see what voters were really seeing and hearing.1 Or for another example, a study mentioned in Chapter 1 resorted to an analysis of newspaper editorials to assign ideology scores to judicial nominees.

- As we will see shortly, case studies provide another alternative. This method involves a researcher examining one or a few individuals, groups, or institutions in great detail. The idea is not to “measure” a few variables but to gain an in-depth understanding of phenomena or to try to understand the world as the subjects do.

Whatever the method, the purpose of nonexperimental designs is to collect information that allows the researcher to approximate the data generated by an experiment and hence make rough causal inferences. To compensate for the inferential shortcomings of the nonexperimental designs it is frequently necessary to achieve a rough approximation of randomization by statistical means. For instance, as demonstrated in later chapters, especially Chapter 13, surveys can gather quantitative and qualitative data, which can then be mathematically manipulated 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 here and in subsequent chapters. In reviewing them, we compare their features to the characteristics of ideal experiments mentioned earlier.

Nonexperimental Time Series Design

Nonexperimental time series designs are characterized by the availability of measures of the dependent variable both before and after the introduction of the independent variable. The researcher does not control the introduction of the independent variable and usually must rely on data collected by others to measure the dependent variable rather than personally conducting the measurements.

In one version of a nonexperimental time series design (sometimes called “interrupted time series analysis”), numerous measurements of a dependent variable are taken both before and after the introduction of the 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 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. These trends may be linear (either increasing or decreasing) or curvilinear, as illustrated in Figure 3-8. After the pre-test trends are established, the researcher then makes several more measurements of the dependent variable after the independent variable has occurred. A change in direction of the measures of the dependent variable away from the existing trend may indicate that the independent variable has had an effect. (In Figure 3-8 such an effect is presumably present in examples B and C, but not in A.) This assumes that nothing else changed that might have affected the dependent variable and that the trend would have continued undisturbed if not for the independent variable.

Time series designs work best when the independent variable occurs at a particular moment or during a fairly brief period of time, affects a dependent variable that is routinely measured, or is known about in advance so that appropriate pre-test measurements can be made. Consequently, this design would work well if we wished to evaluate the impact of a new program or policy initiative. For example, 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

![Figure 3-8](image)

*Examples of Pre-treatment Trends in the Dependent Variable*

A

<table>
<thead>
<tr>
<th>Time</th>
<th>Independent Variable</th>
<th>Dependent Variable</th>
</tr>
</thead>
</table>

B

<table>
<thead>
<tr>
<th>Time</th>
<th>Independent Variable</th>
<th>Dependent Variable</th>
</tr>
</thead>
</table>

C

<table>
<thead>
<tr>
<th>Time</th>
<th>Independent Variable</th>
<th>Dependent Variable</th>
</tr>
</thead>
</table>
Research Design

population, perhaps) may also be changing during the time period under study. If so, we cannot know if it is the checkpoints or the other factor(s) or both that really affect 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. Here we use the word "quasi" to indicate that the researcher does not control the assignment to the experimental or control groups.

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 3-9 we see the hypothetical results of such a time-series-with-quasi-control-group design. We can see that before the introduction of the independent variable the less exposed (quasi-control) group was more supportive of mandatory health care coverage, and that there was already in place 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.

Sometimes a time series analysis consists not only of measures of the dependent variable over an extended period of time but also of measures of the independent variable over a similar period. The challenge here is to look for patterns in both series of measurements that suggest that the independent and dependent variables are related.

A good example of such a study is a recent test of the hypothesis that changes in public opinion affect changes in the character of Supreme Court decisions over time.¹⁷ Researchers William Mishler and Reginald S. Sheehan developed a measure of the overall ideological tenor of Court decisions (the
dependent variable) for each year from 1956 through 1989. They then supplemented that time series with a similar series of measures of the ideological mood of the public each year, as derived from public opinion polls. The results of these two time series over this thirty-three-year period are shown in Figure 3-10.

Although there appear to be similar trends in these two time series, there are many additional questions to be answered about the time lag between the two variables; the alternate manners in which opinion can influence judicial decisions; and the possibility that other factors, such as the political composition of Congress and the ideology of the president, could also influence the ideological nature of Court decisions. Mishler and Sheehan explore all these questions through modifications in their time series design.

Time series studies may be affected by numerous threats to internal validity. For example, instrument change may affect the measurement of the dependent variable over time. This may be a problem if the researcher has

---

**Figure 3-10**

Time Series Analysis of Measures of the Liberalism of the Public Mood and the Liberalism of Supreme Court Decisions

---

relied on existing data collected by others. For example, city crime rates may be compared for several years before and after a change in the way crime statistics are collected has been made. Achievement scores of schoolchildren may be assessed before and after a change in the school's reading program has been made. City budgets may be examined before and after a reorganization of city departments. In these instances, the researcher's conclusions may be jeopardized by the different ways of recording crimes, by the effect of the new reading program on achievement tests, or by different city accounting methods. The longer the time period under study, the more likely that instrument change has occurred.

**Cross-sectional Design**

Another 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, and the researcher does not have any control over the introduction of the independent variable, the assignment of subjects to treatment or control groups, or the conditions under which the independent variable is experienced. In fact, these things are simply observed or recorded, and in the case of surveys the respondents themselves frequently report their exposure to various factors. The measurements are used to construct, with the help of statistical methods, post-treatment quasi-experimental and quasi-control groups that have naturally occurred, 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 then 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 has the virtues of allowing observation of phenomena in more natural, realistic settings; increasing the size and representativeness of the populations studied; and allowing the testing of hypotheses that do not lend themselves to experimental manipulation. In short, cross-sectional research 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 dividing the respondents into different categories according to their answers. To take a slightly more realistic case we might want to test the hypothesis "those with more formal education earn more income" by using a cross-sectional design. We could survey a random sample of adults, ask them questions about their income and educational attainment, and divide them into groups based on their differing levels of
formal education. In this way we observe the quasi-control group (those with no formal education) and the quasi-experimental group (all others). (Notice that we did not control who would be in each group by forcing people to have differing amounts of formal education. The groups were simply naturally occurring and observed.) We could then measure and contrast the income levels of each of these groups to assess the impact of education on income.

If the incomes of those with greater educational attainment were higher than those with less education, we would have accumulated evidence that education and income are related. Because of our research design, however, and our inability to ensure that those with less and those with more education were alike in every other way, we could not necessarily conclude that education determined income. There may be other ways in which those with less and those with more education are different (gender, age, race, geographical location, for example) that also have an effect on income. With a cross-sectional design, we typically employ data analysis techniques to control for variables that may affect both the independent and dependent variables. So if we wanted to control for these factors, we would have to include appropriate questions in the survey and then use statistics to hold them constant.

The cross-sectional design is frequently used in survey research (see Chapter 10). In a study of attitudes toward busing, for example, researchers used survey data to measure the dependent variable (support for busing) and the independent variables (racial intolerance, political conservatism, and self-interest in the busing issue—that is, whether or not the respondent had a child in school) at the same point in time. They divided people into groups based on the measures of the independent variables and then observed the amount of the dependent variable in each of the groups. They found that measures of racial intolerance and political conservatism were inversely correlated with support for busing, while measures of self-interest and support for busing were not related.

This study and others of the same type have several limitations. Because all measures were taken at one point in time, the researchers could not be certain that racial intolerance or political conservatism preceded attitudes toward busing. Therefore, they could not assert that racial intolerance caused attitudes toward busing, even though there was a strong relationship between the two. The respondents' attitudes toward busing, after all, may have been formed before their development of racial intolerance. Furthermore, since the subjects were not randomly assigned to treatment groups, any differences between treatment groups may have resulted from factors other than the independent variables under consideration.

As we noted earlier, 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 Chapter 13. Yet holding these factors constant is problematic, since it is very difficult to be sure that all relevant variables have been explicitly identified and measured. It is important to stress that if a causal variable is not recognized and brought into the analysis, its effects are nonetheless still operative.

Edward Tufte's use of a cross-sectional design in his study of whether compulsory automobile safety inspection programs help reduce traffic fatalities illustrates an attempt to control for variables that might disguise the effect of an independent variable in a cross-sectional study. 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.

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, it may be that there has always been a difference in the auto death rates of these two groups of states and that the existence or absence of state inspections is irrelevant. Even before adopting inspection programs, some states may have had very low death rates for reasons that have nothing to do with car inspections. 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 actually understate the benefits of inspection programs, especially if many of the inspection programs were weak or poorly implemented. This possibility 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. One way of improving
upon the cross-sectional design is by introducing a pre-test of the independent variable.

**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 thus be used to observe changes over time and to provide 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, 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 classic 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 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 classic 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. 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." 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.

Case Study Design
The final nonexperimental research design we will discuss is the case study. In a case study 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. For many years the case study was considered to be an inferior research strategy, but it is now recognized as a "distinctive form of empirical inquiry" and an important design to use for the development and evaluation of public policies as well as for developing explanations for and testing theories of political phenomena.
Research Design

Robert K. Yin, one of the leading proponents of the case study design, defines the case study as an empirical inquiry that (1) investigates a contemporary phenomenon within its real-life context; when (2) the boundaries between phenomenon and context are not clearly evident; and in which (3) multiple sources of evidence are used.\textsuperscript{19} Yin distinguishes between histories and case studies, reserving the term case study for the study of contemporary events.\textsuperscript{20} Other researchers do not make this distinction, but the study of contemporary events does allow researchers a wider selection of data collection methods, including observation and interviewing.\textsuperscript{21}

A case study may be used for exploratory, descriptive, or explanatory purposes. Exploratory case studies may be conducted when little is known about some political phenomenon. Researchers initially may observe only one or a few cases of that phenomenon. Careful observation of a small number of cases may suggest possible general explanations for the behavior or attributes that are observed. These explanations—in the form of hypotheses—can then be tested more systematically by observing more cases. Carefully observing the origins of political dissent within one group at one location may suggest general explanations for dissent, and observing a handful of incumbent representatives when they return to their districts may suggest hypotheses relating incumbent attributes, district settings, and incumbent-constituency relations.\textsuperscript{22}

In the descriptive case, the purpose of a case study may be to find out and describe what happened in a single or select few situations. The emphasis is not on developing general explanations for what happened.

According to Yin, case studies are most appropriately used to answer "how" or "why" questions.\textsuperscript{23} 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. Case studies are particularly useful for testing hypotheses deduced from existing theories of politics.

Proponents argue that the case study 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, is causal.\textsuperscript{24} 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 actually caused the dependent variable. That is, they can learn whether there is an actual link 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.\textsuperscript{25}
The case study design differs from experimental designs in that the researcher is able neither to assign subjects or cases to experimental and control groups nor to manipulate the independent variable. Furthermore, the researcher does not control the context or environment as in a laboratory experiment. Yet the researcher can, through the careful selection of a case or cases, achieve a quasi-experimental situation. For example, a researcher 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. Furthermore, 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 actually 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.

Like experimental and other nonexperimental research designs, the case study design has several variations. In a single case study, the researcher focuses on a single unit of analysis, such as a single group, neighborhood, bureaucracy, or program. In some situations a single case may represent a critical test of a theory.25

An example that demonstrates the explanatory possibilities of a single case study is Jeffrey L. Pressman and Aaron B. Wildavsky's study of the implementation of an economic development program in Oakland, California.27 In contrast to earlier programs that had failed, the Oakland program lacked certain factors associated with failure: a high level of conflict, excessive publicity, political importance and sensitivity, and insufficient funds. Yet the Oakland program also failed. Pressman and Wildavsky attributed the program's failure to the fact that numerous approvals and clearances had to be obtained from a variety of participants. These "perfectly ordinary circumstances" led to the unraveling of previous agreements and ultimately the demise of the Oakland program.28

By choosing a case in which implementation looked as if it would be easy, Pressman and Wildavsky were able to shed considerable light on the process of implementation. This type of case study design has been called the deviant case study, a case that differs from what prevailing theory would lead the researcher to expect. The researcher looks for factors that may explain why the case differs. Research like this may lead to the revision or clarification of existing theories.

Another good example of a deviant case study is discussed in Union Democracy by Seymour Martin Lipset, Martin Trow, and James Coleman.29 It had long been observed that voluntary organizations conform to Robert Michels's "iron law of oligarchy."30 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 adminis-
Research Design

tration, usually retains power indefinitely, rarely faces organized opposition, and when faced with such opposition often resorts to undemocratic procedures to eliminate it. 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 understand this anomaly and in doing so helped explain the workings of democratic processes in general.

Case studies may involve more than one case. A multiple case study is more likely to have explanatory power than a single case study because it provides the opportunity 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. 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.

Despite the important contribution to our understanding of political phenomena a researcher can make with case study research, there are some concerns about the knowledge generated by case studies. One concern about case studies is the "lack of rigor" in presenting evidence and the possibility for bias in the use of evidence. Typically, researchers sift through enormous quantities of detailed information about their cases. In studying contemporary events, the researcher may be the only one to record certain behavior or phenomena. Certainly, the potential for bias is not limited to case studies.

Another frequently raised criticism of case studies is the inability to generalize from a single case. 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. 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)."

A third potential drawback of case studies is that they may take a long time to conduct and result in lengthy reports owing to the need to present adequate documentation to support one's conclusions. 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.
Still, in many circumstances the case study design can be an informative and appropriate research design in many circumstances. 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 actually 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.

**Alternative Research Strategies**

To study a phenomenon like the effects of the media on voters or the behavior of federal justices political scientists most commonly employ 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 totally 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. Aeronautical engineers, for example, use models to see how certain wing shapes affect a plane's stability or what a sudden downdraft will do to its structural integrity. So, even though model planes may be woefully "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 "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.

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; (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