

CHAPTER 4

General Issues in Research Design

Here, we'll examine some fundamental principles about conducting empirical research: causation, and variations on whom or what is to be studied, when, and how. We'll also take a broad overview of the research process.

Learning Objectives

1. Recognize how explanatory scientific research centers on the notion of cause and effect, and why this is a probabilistic model of causation.
2. Describe the three basic requirements for establishing a causal relationship in science, together with what is a necessary cause and a sufficient cause.
3. Understand the role of validity and threats to validity of causal inference.
4. Summarize the four classes of validity threats, and how they correspond to questions about cause and effect.
5. Discuss how a scientific realist approach bridges idiographic and nomothetic approaches to causation.
6. Describe different units of analysis in criminal justice research.
7. Explain how the ecological fallacy relates to units of analysis.
8. Understand the time dimension, together with the differences between cross-sectional and longitudinal research.
9. Describe how retrospective studies may approximate longitudinal studies.

Introduction 84

Causation in the Social Sciences 84

- Criteria for Causality 85
- Necessary and Sufficient Causes 86

Validity and Causal Inference 87

- Statistical Conclusion Validity 87
- Internal Validity 88
- External Validity 88
- Construct Validity 89
- Validity and Causal Inference Summarized 90
- Does Drug Use Cause Crime? 90
- Introducing Scientific Realism 92

Units of Analysis 92

- Individuals 94
- Groups 94
- Organizations 94

Parts-Marking and Auto Theft

Ronald Clarke and Mike Maxfield conducted a study to examine the effect of parts-marking on auto theft. Their work was commissioned by the National Highway Traffic Safety Administration (NHTSA). Beginning in 1987, NHTSA required that certain body and other parts be marked with vehicle identification numbers (VIN) on cars subject to high rates of theft. Though the question "Does parts-marking produce a decline in auto theft?" seemed straightforward at first, Clarke and Maxfield began by breaking that general question into a series of research questions:

- Has the decline in auto theft since 1987 been greater than the decline in other crimes?
- What have been the trends in thefts from cars? Have they been different than theft of cars?
- Which cars are required to have parts marked?
- Has there been a greater decline in thefts of cars that are required to have marked parts compared to those where parts-marking is not required?

CHAPTER 4 General Issues in Research Design 83

- Social Artifacts 95
- The Ecological Fallacy 95
- Reductionism 96

Units of Analysis in Review 97

- UNIT OF ANALYSIS IN THE NATIONAL YOUTH GANG SURVEY 98

The Time Dimension 98

- Cross-Sectional Studies 99
- Longitudinal Studies 100
- Approximating Longitudinal Studies 101
- Retrospective and Prospective Studies 101
- The Time Dimension Summarized 103

PUTTING IT ALL TOGETHER: CAUSE, UNITS, AND TIME 104

Answers to the Units-of-Analysis Exercise 106

Summary 106

- Which cars have other forms of security, such as transponder keys and electronic immobilizers?
- How does the trend in cars equipped with other forms of security compare to trends in cars with parts-marking?

These are a few of the several questions that were formulated to guide a research project that became considerably more complex than Clarke and Maxfield initially thought. Their work illustrates many of the topics covered in this chapter. Asking whether car theft declined over time more than other crimes involved the time dimension and a basic question about causation. If car theft had not declined more than other crimes, it was unlikely that federal requirements for parts-marking could have caused a decline.

Different units of analysis were involved. Clarke and Maxfield were basically interested in aggregate trends of car theft over a long period of time. But they had to determine which individual cars fit into different categories,

which required examining model years, such as 2009 Ford Mustangs equipped with 5-liter engines. These and other general issues in research design were initially addressed at early stages of the project. But as Clarke and Marfield learned more, they had to adjust their approach and develop new measures. That's an example of our opening statement to this part of the book: posing questions properly is often more difficult than answering them. It is common for researchers to make adjustments as they move from planning to executing a particular research project. Review this box as you read the chapter. Consider another example of a seemingly simple question that can become more complex: Has the growth of online shopping contributed to an increase in identity theft? Or think of another example of a seemingly simple research question that becomes more complex as you begin to think through it.

Introduction

Causation, units, and time are key elements in planning a research study.

Science is an enterprise dedicated to "finding out." No matter what we want to find out, though, there are likely to be a great many ways of doing so. Topics examined in this chapter address important considerations in planning scientific inquiry—how to design a strategy for finding out something. Often, criminal justice researchers want to find out something that involves questions of cause and effect. They may want to find out more about things that make crime more likely to occur or about policies that they hope will reduce crime in some way.

Let's say you are interested in studying corruption in government. That's certainly a worthy and appropriate topic for criminal justice research. But what specifically are you interested in? What do you mean by "corruption"? What kinds of behavior do you have in mind? And what do you mean by "government"? Whom do you want to study: public employees? Judges? Sworn police officers? Elected officials? What is your purpose? Do

you want to find out how much corruption there is? Do you want to learn why corruption exists? Do you want to compare corruption in different agencies? In different cities or states? These are the kinds of questions that need to be answered in research design. Read the vignette titled "Parties, Marking and Auto Theft" for an example of how a single question can lead to many more.

In practice, all aspects of research design are interrelated. They are separated here and in subsequent chapters so that we can explore particular topics in detail. We start with a discussion of causation in social science, the foundation of explanatory research. We then examine units of analysis—the *what* or *whom* to study. Deciding on units of analysis is an important part of all research, partly because people sometimes inappropriately use data measuring one type of unit to say something about a different type of unit.

Next, we consider alternative ways of handling time in criminal justice research. It is sometimes appropriate to examine a static cross section of social life, but other studies follow social processes over time. In this regard, researchers must consider the time order of events and processes in making statements about cause.

Causation in the Social Sciences

Causation is the focus of explanatory research.

Cause and effect are implicit in much of what we have examined so far. One of the chief goals of social science researchers is to explain why things are the way they are. Typically, we do that by specifying the causes for the way things are: some things are caused by other things.

The general notion of causation is both simple and complex. On the one hand, we could have ignored the issue altogether. We could simply plug in the terms *cause* and *effect* and have little difficulty understanding them. On the other hand, an adequate discourse on causation would require a whole book or even a series of books. We decided to adopt a middle ground, providing more than

a commonsense perspective on causation but not attempting to be definitive.

Much of our discussion in this section describes issues of causation and validity for social science in general. Recall from Chapters 1 and 2 that criminal justice research and theory are most strongly rooted in the social sciences. Furthermore, social science research methods are adapted from those used in the physical sciences. Many important and difficult questions about causality and validity occupy researchers in criminal justice, but our basic approach requires stepping back a bit to consider the larger picture of how we can or cannot assert that some cause actually produces some effect.

At the outset, it's important to keep in mind that cause in social science is inherently **probabilistic**. As a point we introduced in Chapter 2, We say, for example, that certain factors make delinquency more or less *likely* within groups of people. Thus, victims of childhood abuse or neglect are *more likely* to report alcohol abuse as adults (Schuck and Widom, 2001). Burglary is *less likely* in neighborhoods where opportunities for informal guardianship are greater (Reynald, 2011).

Criteria for Causality

We begin our consideration of cause by examining what criteria must be satisfied before we can infer that something causes something else. Recall the discussion of idiographic and nomothetic modes of explanation in Chapter 2. We now consider, first, criteria for inferring idiographic causation, then criteria for inferring cause in the nomothetic mode of explanation.

Joseph Maxwell (2005:106–107) writes that criteria for assessing an idiographic explanation are (1) how credible and believable it is and (2) whether alternative explanations ("rival hypotheses") were seriously considered and found wanting. The first criterion relates to logic as one of the foundations of science. We demand that our explanations make sense, even if the logic is sometimes complex. The second criterion reminds us of Sherlock Holmes's dictum that when all other possibilities have been eliminated, the remaining explanation, however improbable, must be the truth.

Regarding nomothetic explanation, we examine three specific criteria for causality, as described by William Shadish, Thomas Cook, and Donald Campbell (2002):

1. Two variables must vary together—they must be empirically correlated.
2. The cause must occur before the effect.
3. The empirical correlation between cause and effect is not due to some other factor.

The first requirement in a causal relationship between two variables is that the two variables be empirically correlated with each other—they must occur together. It makes no sense to say that exploding gunpowder causes a bullet to leave the muzzle of a gun if, in observed reality, a bullet does not come out after the gunpowder explodes.

As straightforward as it may seem, criminal justice research has difficulties with this requirement. In the probabilistic world of nomothetic models of explanation, at least, we encounter few perfect measures of correlation. Most judges sentence repeat drug dealers to prison, but some don't. Many gang members use illegal drugs, but many do not. We are forced to ask, therefore, how strong the empirical relationship must be for that relationship to be considered causal.

The second requirement for inferring a causal relationship between two variables is that the cause precede the effect in time. It makes no sense to imagine something being caused by something else that happened later on. A bullet leaving the muzzle of a gun does not cause the gunpowder to explode; it works the other way around.

As simple and obvious as this criterion may seem, criminal justice research suffers many problems in this regard. Often, the time order connecting two variables is simply unclear. Which comes first: drug use or crime? And even when the time order seems clear, exceptions may be found. For example, we normally assume that obtaining a master's degree in management is a cause of more rapid advancement in a state department of corrections. Yet corrections executives might pursue graduate education after they have been promoted and recognize that advanced training in management skills will help them do their job better.

The third requirement for a causal relationship is that the observed empirical correlation between two variables cannot be explained away as being due to the influence of some third variable that causes both of them. For example, we may observe that drug markets are often found near bus stops, but this does not mean that bus stops encourage drug markets. A third variable is at work here: groups of people naturally congregate in reasonably close proximity at bus stops, and street drug markets are often found where people naturally congregate. In fact, Sung-suk Violet Yu (2011) describes a number of variables that are involved in a complex relationship between bus stops and crime.

To sum up, most criminal justice researchers consider two variables to be causally related—one causes the other—if (1) there is an empirical correlation between them, (2) the cause precedes the effect in time, and (3) the relationship is not found to result from the effects of some third variable on each of the two initially observed. Any relationship that satisfies all these criteria is causal.

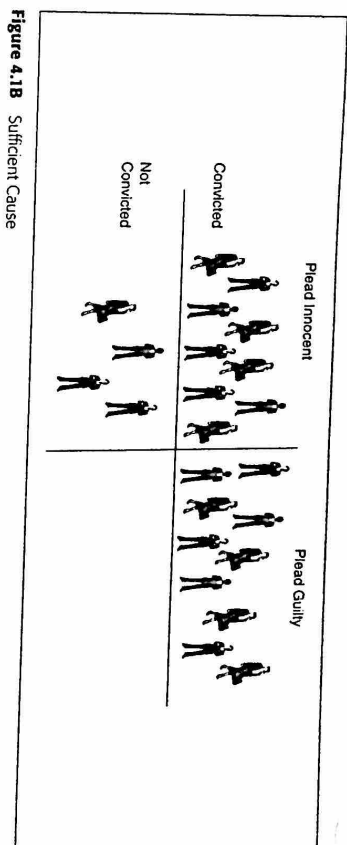
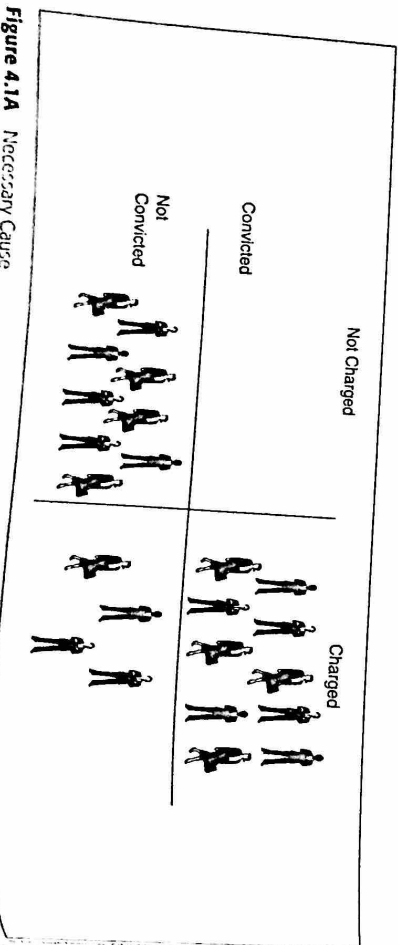
Necessary and Sufficient Causes

Recognizing that virtually all causal relationships in criminal justice are probabilistic is central to understanding other points about cause. Within the probabilistic model, it is useful to distinguish

two types of causes: necessary and sufficient causes. A *necessary* cause is a condition that must be present for the effect to follow. For example, it is necessary for someone to be charged with a criminal offense before a person can be convicted. Being charged is not sufficient. Some people who are charged are ruled innocent. You must be guilty or be found guilty by the court. Figure 4.1a illustrates a necessary cause—it is necessary to be charged before one can be convicted.

A *sufficient* cause, in contrast, is a condition that more or less guarantees the effect in question. Thus, for example, pleading guilty to some criminal charge is a sufficient cause for being convicted. It is also possible to be convicted through a trial as well, but pleading guilty is sufficient. Figure 4.1b illustrates this state of affairs.

The discovery of a cause that is both necessary and sufficient to produce some result is the most satisfying outcome in research. If we are studying juvenile delinquency, we want to discover a single condition that (1) has to be present for delinquency to develop and (2) always results in delinquency. Then we will surely feel that we know precisely what causes juvenile delinquency. Unfortunately, we seldom discover causes that are both necessary and sufficient. Most causal relationships that criminal justice researchers work with are probabilistic and partial—we are able to partly explain cause and effect in some percentage of cases we observe.



Validity and Causal Inference

Scientists assess the truth of statements about cause by considering threats to validity.

Paying careful attention to cause-and-effect relationships is crucial in criminal justice research. Cause and effect are also key elements of applied studies, in which a researcher may be interested in whether, for example, a new mandatory sentencing law actually causes an increase in the prison population.

When we are concerned with whether we are correct in inferring that a cause produced an effect, we are concerned with the validity of causal inference. In the words of Shadish, Cook, and Campbell (2002:34), validity is “the approximate truth of an inference.... When we say something is valid, we make a judgment about the extent to which relevant evidence supports that inference as being true or correct.” They emphasize that *approximate* is an important word because one can never be absolutely certain about cause.

It is almost never possible to conclusively establish the validity of causal inference. We cannot make any sort of absolute judgment about whether our statements about cause are true. Instead, social scientists try to address different validity threats in causal inference—reasons

we might be incorrect in stating that some cause produced some effect. As Maxwell (2005:106) puts it, “A key concept for validity is thus the *validity* threat: a way you might be wrong” (emphasis in original). Here we will summarize the threats to four general categories of validity: statistical conclusion validity, internal validity, construct validity, and external validity. Chapter 7 discusses each type in more detail, linking the issue of validity to different ways of designing research.

Statistical Conclusion Validity

Statistical conclusion validity refers to our ability to determine whether a change in the suspected cause is statistically associated with a change in the suspected effect. This corresponds with the first criterion for inferring cause: Are two variables related to each other? If we suspect that using illegal drugs causes people to commit crimes, one of the first things we will be interested in is the common variation between drug use and crime. If drug users and nonusers commit equal rates of crime and if about the same proportions of

Validity Whether statements about cause or measures are correct (valid) or false (invalid).

Validity threats Possible sources of false conclusions about cause or measurement.

criminals and noncriminals use drugs, there will be no statistical relationship between measures of drug use and criminal offending. That seems to be the end of our investigation of the causal relationship between drugs and crime. Review the vignette at the opening of this chapter for another example, the first research question considered by Clarke and Maxfield.

Basing conclusions on a small number of cases is a common threat to statistical conclusion validity. For example, suppose a researcher studies ten drug users and ten nonusers and compares the numbers of times these subjects are arrested for other crimes over a 6-month period. The researcher might find that the ten users were arrested an average of three times while nonusers averaged two arrests in 6 months. There is a difference in arrest rates, but is it a significant difference? Statistically, the answer is no because so few drug users were included in the study. Researchers cannot have much confidence in statements about cause if their findings are based on a small number of cases.

Threats to statistical conclusion validity might also have the opposite effect of suggesting that covariation is present when, in fact, there is no cause-and-effect relationship. The reasons for this are again somewhat technical and require a basic understanding of statistical inference. However, some of the superstitious behavior exhibited by gamblers provides a rough-and-ready example.

Let's say you like to play the state lottery; nothing serious, you simply buy a couple of tickets each week. Many people play a system whereby they pick numbers that correspond to their date of birth or their license plate number. Expecting these lucky numbers to influence your chances of winning the lottery is an example of failing to recognize that winning numbers are picked by a random process. If you always pick your favorite uncle's birthday—say, July 7, 1977 (7777)—and buy two tickets per week for 5 years, your lucky number might come up simply by chance in the random system used to pick winning numbers. But there is no statistical conclusion validity to the inference that consistently picking your birthday will cause you to win the lottery. Your winning will be due to the random effects of how numbers are selected.

Internal Validity

Even if we have found that two variables appear to be related to each other, **internal validity** threats challenge causal statements about the observed covariation. An observed association between two variables has internal validity if the relationship is, in fact, causal and not due to the effects of one or more other variables. Whereas statistical conclusion threats are most often due to random error, internal validity problems result from nonrandom or systematic error. Simply put, threats to the internal validity of a proposed causal relationship between two variables usually arise from the effects of one or more other variables. Notice how this validity threat relates to the third requirement for establishing a causal relationship: eliminating other possible explanations for the observed relationship.

If, for example, we observe that convicted drug users sentenced to probation are rearrested less often than drug users sentenced to prison, we might be tempted to infer that prison sentences cause recidivism. Although being in prison might have some impact on whether someone commits more crimes in the future, in this case it is important to look for other causes of recidivism. One likely candidate would be prior criminal record. Convicted drug users without prior criminal records are more likely to be sentenced to probation, whereas persons with previous convictions more often receive prison terms. Research on criminal careers has found that the probability of reoffending increases with the number of prior arrests and convictions (Blumstein et al. 2010). In this case, a third variable—prior convictions—may explain some or all of the observed tendency of prison sentences to be associated with recidivism. Prior convictions are associated with both sentence—prison or probation—and subsequent convictions.

External Validity

Are findings about the impact of mandatory arrest for family violence in Minneapolis similar to findings in Milwaukee? Can community crime prevention organizations successfully combat drug use throughout a city, or do they work best in areas with only minor drug problems? Electronic

monitoring may be suitable as an alternative sentence for convicted offenders, but can it work as an alternative to jail for defendants awaiting trial? Such questions are examples of issues in **external validity**. Do research findings about cause and effect apply equally to different cities, neighborhoods, and populations?

In a general sense, external validity is concerned with whether research findings from one study can be reproduced in another study, often under different conditions. Because crime problems and criminal justice responses can vary so much from city to city or from state to state, researchers and public officials are often especially interested in external validity. For example, a Kansas City evaluation found sharp reductions in gun-related crimes in hot spots that had been targeted for focused police patrols (Sherman, Shaw, and Rogan, 1995). Because these results were promising, similar projects were launched in two other cities—Indianapolis (McGarrell et al., 2001) and Pittsburgh (Cohen and Ludwig, 2003). In both cases, researchers found that police actions targeting hot spots for gun violence reduced gun-related crime and increased seizures of illegal firearms. Having similar findings in Indianapolis and Pittsburgh enhanced the external validity of original results from Kansas City.

Construct Validity

This type of validity is concerned with how well an observed relationship between variables that a researcher has measured represents the underlying causal process of interest. In this sense, **construct validity** refers to generalizing from what we observe and measure to the real-world things in which we are interested. The concept of construct validity is thus closely related to issues in measurement, as we will see in Chapters 5 and 6.

To illustrate construct validity, let's consider the supervision of police officers—specifically, whether close supervision causes police officers to write more traffic tickets. We might define “close supervision” in this way: A police sergeant drives his own marked police car in such a way as to always keep a patrol car in view.

This certainly qualifies as close supervision, but you may recognize a couple of problems. First, two marked patrol cars present a highly visible presence to motorists, who might drive more prudently and thus reduce the opportunities for patrol officers to write traffic tickets. Second, and more central to the issue of construct validity, this represents a narrow definition of the construct “close supervision.” Patrol officers may be closely supervised in other ways that cause them to write more traffic tickets. For example, sergeants might closely supervise their officers by reviewing their ticket production at the end of each shift. Supervising subordinates by keeping them in view is only one way of exercising control over their behavior. It may be appropriate for factory workers, but it is not practical for police, representing a very limited version of the construct “supervision.”

The well-known Kansas City Preventive Patrol Experiment, discussed in Chapter 1, provides another example of construct validity problems. Recall that the experiment sought to determine whether routine preventive patrol caused reductions in crime and fear of crime and increases in arrests. This causal proposition was tested by comparing measures of crime, fear, and arrests in proactive beats (with twice the normal level of preventive patrol), reactive beats (with no preventive patrol), and control beats (with normal levels of preventive patrol). Researchers found no significant differences in levels of crime, fear, and arrests.

Richard Larson (1973) discussed several difficulties with the experiment's design. One important problem relates to the visibility of police presence, a central concept in preventive patrol. It is safe to assume that the ability of routine patrol to prevent crime and enhance feelings of safety depends crucially on the visibility of police. It makes sense to assume further that withdrawing preventive patrol, as was done in the reactive beats, reduces the visibility of police. But by how much? Larson explored this question in detail and suggested that two other features of police operations during the Kansas City experiment partially compensated for the absence of preventive patrol and produced a visible police presence.

First, the different types of experimental beats were adjacent to one another; one reactive beat

shared borders with three control and three proactive beats. This enhanced the visibility of police in reactive beats in two ways: (1) police in adjoining proactive and control beats sometimes drove around the perimeter of reactive beats, and (2) police often drove through reactive beats on their way to some other part of the city.

Second, many Kansas City police officers were skeptical about the experiment and feared that withdrawing preventive patrol in reactive beats would create problems. As a result, police who responded to calls for service in the reactive areas more frequently used lights and sirens when driving to the location of complaints. A related action was that police units not assigned to the calls for service nevertheless drove into the reactive beats to provide backup service.

Each of these actions produced a visible police presence in the reactive beats. People who lived in these areas were unaware of the experiment and, as you might expect, did not know whether a police car happened to be present because it was on routine patrol, was on its way to some other part of the city, or was responding to a call for assistance. And, of course, the use of lights and sirens makes police cars much more visible.

Larson's point was that the *construct* of police visibility is only partly represented by routine preventive patrol. A visible police presence was produced in Kansas City through other means. Therefore, the researchers' conclusion that routine preventive patrol does not cause a reduction in crime or an increase in arrests suffers from threats to construct validity. Construct validity is a frequent problem in applied studies, in which researchers may oversimplify complex policies and policy goals. That is basically a measurement problem, as we will see in the next chapter.

Validity and Causal Inference Summarized

The four types of validity threats can be grouped into two categories: bias and generalizability. Internal and statistical conclusion validity threats are related to systematic and nonsystematic bias, respectively. Problems with statistical procedures produce nonsystematic bias, while an alternative explanation for an observed relationship is an

example of systematic bias. In either case, bias calls into question the inference that some cause produced some effect. Failing to consider the more general cause-and-effect constructs that operate in an observed cause-and-effect relationship results in research findings that cannot be generalized to real-world behaviors and conditions. And a cause-and-effect relationship observed in one setting or at one time may not operate in the same way in a different setting or at a different time.

Shadish, Cook, and Campbell (2002:39) summarized their discussion of these four validity threats by linking them to the types of questions that researchers ask in trying to establish cause and effect. Test your understanding by writing the name of each validity threat after the appropriate question from Shadish, Cook, and Campbell.

1. How large and reliable is the covariation between the presumed cause and effect? _____
2. Is the covariation causal, or would the same covariation have been obtained without the treatment? _____
3. How generalizable is the locally embedded causal relationship over varied persons, treatments, observations, and settings? _____
4. What general constructs are involved in the persons, settings, treatments, and observations used in the experiment? _____

Does Drug Use Cause Crime?

As a way of illustrating issues of validity and causal inference, we will consider the relationship between drug use and crime. Drug addiction is thought to cause people who are desperate for a fix and unable to secure legitimate income to commit crimes in order to support their habit.

Understanding causal statements about drug use and crime requires carefully specifying two key concepts—drug use and crime—and considering the different ways these concepts might be related. Jan Chaiken and Marcia Chaiken (1990) provide unusually careful and well-reasoned insights that will guide our consideration of links

between drugs and crime. Let's begin with issues in basic research on the drugs-crime connection and then discuss the implications of this research for criminal justice policy.

First is the question of temporal order: Which comes first, drug use or crime? Research summarized by Chaiken and Chaiken provides no conclusive answer. In an earlier study of prison inmates, Chaiken and Chaiken (1982) found that 12 percent of their adult subjects committed crimes after using drugs for at least 2 years, while 15 percent committed predatory crimes 2 or more years before using drugs.

Regarding an empirical relationship, many studies have found that some drug users commit crimes and that some criminals use drugs. Trevor Bennett and associates (2008) conducted a summary analysis of 30 research studies on drug use and crime. They found that drug users were over three times more likely to be involved in crime than nonusers. So a statistical association between drug use and crime clearly exists.

What about the possible influence of other factors, things that might be related to both drug use and crime? Chaiken and Chaiken (1990:234) conclude that "drug use and crime participation are weakly related as contemporaneous products of factors generally antithetical to traditional United States lifestyles." Stated somewhat differently, drug use and crime (as well as delinquency) are each deviant activities produced by other underlying causes. A statistical association between drug use and crime clearly exists. But the presence of other factors indicates that the relationship is not directly causal, thus bringing into question the internal validity of causal statements about drug use and crime.

To assess the construct validity of research on drugs and crime, let's think for a moment about different patterns of each behavior, rather than assume that drug use and crime are uniform behaviors. Many adolescents in the United States experiment with drugs, and just as many—especially males—commit delinquent acts or petty crimes. A large number of adults may be occasional users of illegal drugs as well. Many different patterns of drug use, delinquency, and adult criminality have been found in research in other

countries. Stephen Pudney (2002) and Trevor Bennett and Kay Holloway (2003) report on varying patterns of use in England and Wales. Because there is no simple way to describe either construct, searching for a single cause-and-effect relationship misrepresents a complex causal process.

Both construct and external validity are concerned with generalizations. Problems with the external validity of research on drugs and crime are similar to those revolving around construct validity. The relationship between occasional marijuana use and delinquency among teenagers is different from that between occasional cocaine use and adult crime; each relationship, in turn, varies from that between heroin addiction and persistent criminal behavior among adults.

The issue of external validity comes into sharper focus when we shift from basic research that seeks to uncover fundamental causal relationships to criminal justice policy. Chaiken and Chaiken argue that any uniform policy to reduce the use of all drugs among all population groups will have little effect on serious crime. Bennett and associates reach similar conclusions, arguing that government antidrug policy should distinguish more carefully between types of drug use (2008:117–118).

Basic and applied research on the relationships among drug use and crime readily illustrates threats to the validity of causal inference. It is often difficult to find a relationship because there is so much variation in drug use and crime participation (statistical conclusion validity threat). A large number of studies have demonstrated that, when statistical relationships are found, both drug use and crime can be attributed to other, often multiple, causes (internal validity threat). Different patterns among different population groups mean that there are no readily identifiable cause-and-effect constructs (construct validity). Because of these differences, policies developed to counter drug use among the population as a whole cannot be expected to have much of an impact on serious crime (external validity).

None of the above is to say that there is no cause-and-effect relationship between drug use and crime. However, research has clearly shown that there is no simple causal connection.

Introducing Scientific Realism

In our final consideration of cause and effect in this chapter, we revisit the distinction between idiographic and nomothetic ways of explanation. Doing research to find what causes what most often reflects nomothetic concerns. We wish to find causal explanations that apply generally to situations beyond those we actually study in our research. At the same time, researchers and public officials are often interested in understanding specific causal mechanisms in more narrowly defined situations—what we have described as the idiographic mode of explanation.

Scientific realism bridges idiographic and nomothetic approaches to explanation by seeking to understand how causal *mechanisms* operate in specific *contexts*. Traditional approaches to finding cause and effect usually try to isolate causal mechanisms from other possible influences, something you should now recognize as trying to control threats to internal validity. The scientific realist approach views these other possible influences as contexts in which causal mechanisms operate. Rather than try to exclude or otherwise control possible outside influences, scientific realism studies how such influences are involved in cause-and-effect relationships.

For example, earlier in this chapter, we noted that electronic monitoring as a condition of probation might apply to some populations but not others. We framed this as a question of external validity in the traditional way of considering nomothetic causation. A scientific realist approach would consider the causal mechanism underlying electronic monitoring to be effective in some contexts but not in others. As another example, we reviewed at some length the cause-and-effect continuum surrounding drug use and crime. That review was framed by traditional nomothetic research to establish cause and effect. A scientific realism approach to the question would recognize that drug use and crime co-occur in some contexts but not in others.

Units of analysis The things—what or whom—being studied in a research project.

We say that scientific realism bridges idiographic and nomothetic modes of explanation because it exhibits elements of both. Because it focuses our attention on very specific questions, scientific realism seems idiographic. Nicholas Zarin, Jon Shane, and Ronald Clarke (2004) asked the question, “Will redesigning the Interstate 78 exit in Newark, New Jersey, cause a reduction in suburban residents seeking to buy heroin in a nearby neighborhood?” But their analysis and interpretation addressed a more general question of causation: “Can the design of streets and intersections be modified to make it more difficult for street drug markets to operate?” Changing an expressway exit ramp to reduce drug sales in Newark is a specific example of cause and effect that is rooted in the more general causal relationship between traffic patterns and drug markets.

These illustrations of the scientific realist approach to cause and effect are examples of research for the purpose of application, a topic treated at length by British researchers Ray Pawson and Nick Tilley (1997). Application is a type of explanatory research, as we indicated in Chapter 1. In later chapters, we call on scientific realism as a strategy for designing explanatory research (Chapter 7) and conducting evaluations (Chapter 13).

Sorting out causes and effects is one of the most difficult challenges of explanatory research. Our attention now turns to two other important considerations that emerge in research for explanation and other purposes: units of analysis and the time dimension.

Units of Analysis

To avoid mistaken inferences, researchers must carefully specify the people or phenomena that will be studied.

In criminal justice research, there is a great deal of variation in what or who is studied—what are technically called **units of analysis**. Individual people are often units of analysis. Researchers may make observations describing certain characteristics of offenders or crime victims, such as age, gender,

or race. The descriptions of many individuals are then combined to provide a picture of the population that comprises those individuals.

For example, we may note the age and gender of persons convicted of drunk driving in Fort Lauderdale over a certain period. Aggregating these observations, we might characterize drunk-driving offenders as 72 percent men and 28 percent women, with an average age of 26.4 years. This is a descriptive analysis of convicted drunk drivers in Fort Lauderdale. Although the description applies to the group of drunk drivers as a whole, it is based on the characteristics of individual people convicted of drunk driving.

The same situation could exist in an evaluation study. Suppose we wish to determine whether an alcohol education program reduces repeat arrests of first-time drunk-driving offenders. First, we administer the education program to half the persons convicted of drunk driving in Fort Lauderdale over a period of 2 months; the other half does not receive the program. Next, we observe drunk-driving arrest records for 12 months, keeping track of people who were previously convicted and of whether they received the alcohol education program. Combining our observations, we find that 5 percent of those who received the program were arrested again and that 20 percent of those who did not receive alcohol education were rearrested within a year. The purpose of the study is to evaluate a program, but individual drunk drivers are still the units of analysis.

Units of analysis in a study are typically also the units of observation. Thus, to study what steps people take to protect their homes from burglary, we might observe, perhaps through interviews, individual household residents. Sometimes, however, we “observe” units of analysis indirectly. For example, we might ask individuals about crime prevention measures for the purpose of describing households. We might want to find out whether homes with double-cylinder deadbolt locks are burglarized less often than homes with less substantial protection. In this case, our units of analysis are households, but the units of observation are individual household members who are asked to describe burglaries and home protection to interviewers.

Some studies try to describe or explain more than one unit of analysis. In these cases, the researcher must anticipate what conclusions she or he wishes to draw with regard to which units of analysis. For example, we may want to discover what kinds of college students (individuals) are most successful in their careers; we may also want to learn what kinds of colleges (organizations) produce graduates who are the most successful.

Here’s an example that illustrates the complexity of units of analysis. Murders usually involve one individual who kills another individual. However, when Charis Kubrin and Ronald Weitzer (2003:157) ask, “Why do these neighborhoods generate high homicide rates?” the unit of analysis in that phrase is neighborhood. You can probably imagine some kinds of neighborhoods (e.g., poor, urban) that would have high homicide rates and some (e.g., wealthy, suburban) that would have low rates. In this particular conversation, the unit of analysis (neighborhood) would be categorized in terms of variables such as economic level, locale, and homicide rate.

In their analysis, however, Kubrin and Weitzer were also interested in different types of homicide: in particular, those that occurred in retaliation for some earlier event, such as an assault or insult. Can you identify the unit of analysis common to all of the following excerpts (Kubrin and Weitzer, 2003:163)?

1. “The sample of killings...”
2. “The coding instrument includes over 80 items related to the homicide.”
3. “Of the 2,161 homicides that occurred from 1985 [to] 1995...”
4. “Of those with an identified motive, 19.5 percent (n = 337) are retaliatory.”

In each of these excerpts, the unit of analysis is *homicide* (also called *killings*). Sometimes you can identify the unit of analysis in the description of the sampling methods, as in the first excerpt. A discussion of classification methods might also identify the unit of analysis, as in the second excerpt (80 ways to code the homicides). Often, numerical summaries point the way: 2,161 homicides; 19.5 percent (of the homicides).

To explore this topic in more depth, let's consider several common units of analysis in criminal justice research.

Individuals

Any variety of individuals may be the units of analysis in criminal justice research. This point is more important than it may initially seem. The norm of generalized understanding in social science should suggest that scientific findings are most valuable when they apply to all kinds of people. In practice, however, researchers seldom study all kinds of people. At the very least, studies are typically limited to people who live in a single country, although some comparative studies stretch across national boundaries.

As the units of analysis, individuals may be considered in the context of their membership in different groups. Examples of circumscribed groups whose members may be units of analysis at the individual level are *police*, *victims*, *defendants in criminal court*, *correctional inmates*, *gang members*, and *active burglars*. Note that each of these terms implies some population of individual persons. Descriptive studies having individuals as their units of analysis typically aim to describe the population that comprises those individuals.

Groups

A variety of groups may also be the units of analysis for criminal justice research. This is not the same as studying the individuals within a group. If we study the members of a juvenile gang to learn about teenagers who join gangs, the individual (teen gang member) is the unit of analysis. But if we study all the juvenile gangs in a city to learn the differences between big gangs and small ones, between gangs selling drugs and gangs stealing cars, and so forth, the unit of analysis is the social group (gang).

Police beats or patrol districts might be the units of analysis in a study. A police beat can be described in terms of the total number of people who live within its boundaries, total street mileage, annual crime reports, and whether the beat includes a special facility such as a park or high school. We can then determine, for example, whether beats that

include a park report more assaults than beats without such facilities or whether auto thefts are more common in beats with more street mileage. Here, the individual police beat is the unit of analysis.

Other examples of units at the group level are *households*, *city blocks*, *census tracts*, *cities*, *counties*, and *other geographic regions*. Each of these terms also implies some population of groups. *Street gang* implies some population that includes all street gangs. The population of street gangs could be described, say, in terms of its geographic distribution throughout a city. An explanatory study of street gangs might discover, for example, whether large gangs are more likely than small ones to engage in intergang warfare.

Organizations

Formal political or social organizations may also be the units of analysis in criminal justice research. An example is correctional facilities, which implies, of course, a population of all correctional facilities. Individual facilities might be characterized in terms of their number of employees, status as state or federal prisons, security classification, percentage of inmates who are from racial or ethnic minority groups, types of offenses for which inmates are sentenced to each facility, average length of sentence served, and so forth. We might determine whether federal prisons house a larger or smaller percentage of offenders sentenced for white-collar crimes than do state prisons. Other examples of formal organizations suitable as units of analysis are police departments, courthouses, probation offices, drug treatment facilities, and victim services agencies.

When social groups or formal organizations are the units of analysis, their characteristics are often derived from the characteristics of their individual members. Thus, a correctional facility might be described in terms of the inmates it houses—gender distribution, average sentence length, ethnicity, and so on. In a descriptive study, we might be interested in the percentage of institutions housing only females. Or, in an explanatory study, we might determine whether institutions housing both males and females report, on the average, fewer or more assaults by

inmates on staff compared with male-only institutions. In each example, the correctional facility is the unit of analysis. In contrast, if we ask whether male or female inmates are more often involved in assaults on staff, then the individual inmate is the unit of analysis.

Some studies involve descriptions or explanations of more than one unit of analysis. Consider, for example, an evaluation of community policing programs in selected neighborhoods of a large city. In such an evaluation, we might be interested in how citizens feel about the program (individually), whether arrests increased in neighborhoods with the new program compared with those without it (groups), and whether the police department's budget increased more than the budget in a similar city (organizations). In such cases, it is imperative that researchers anticipate what conclusions they wish to draw with regard to what units of analysis.

Social Artifacts

Yet another potential unit of analysis may be referred to as "social artifacts," or the products of social beings and their behavior. One class of social artifacts is stories about crime in newspapers and magazines or on television. A newspaper story might be characterized by its length, placement on front or later pages, size of headlines, and presence of photographs. A researcher could analyze whether television news features or newspaper reports provide the most details about a new police program to increase drug arrests.

Social interactions are also examples of social artifacts suitable for criminal justice research. Police crime reports are an example. We might analyze assault reports to find how many involved three or more people, whether assaults involve strangers or people with some prior acquaintance, or whether they more often occurred in public or private locations.

At first, crime reports may not seem to be social artifacts, but consider for a moment what they represent. When a crime is reported to the police, officers usually record what happened from descriptions by victims or witnesses. For instance, an assault victim may describe how he suffered

an unprovoked attack while innocently enjoying a cold beer after work. However, witnesses to the incident might claim that the "victim" started the fight by insulting the "offender." The responding police officer must interpret who is telling the truth in trying to sort out the circumstances of a violent social interaction. The officer's report becomes a social artifact that represents one among the population of all assaults.

Records of different types of social interactions are common units of analysis in criminal justice research. Criminal history records, meetings of community antiracket groups, presence investigations, and interactions between police and citizens are examples. Notice that each example requires information about individuals but that social interactions between people are the units of analysis.

In his analysis of juvenile arrests, Howard Snyder (2005) offers examples of how social artifacts sometimes stand in for individuals as units of analysis. Police record what Snyder terms an arrest event. Each arrest event includes at least one person, but because juveniles often commit crimes in groups, a single arrest event might include a large number of juveniles. Furthermore, individual juveniles might be arrested more than once over some time period, so the number of arrests is not the same as the number of juveniles arrested.

The Ecological Fallacy

At this point, it is appropriate to introduce two important concepts related to units of analysis and causation: the ecological fallacy and reductionism. The first concept—the **ecological fallacy**—refers to the danger of making assertions about individuals as the unit of analysis based on the examination of groups or other aggregations. Such assertions are often made in connection with causation, where researchers observe associations between aggregate units and make statements about causality between individual units.

Suppose we are interested in learning about robbery in different police precincts of a large city. Let's assume that we have information on how many robberies were committed in each police precinct of Chicago for the year 2012. Assume also

that we have census data describing some of the characteristics of those precincts. Our analysis of such data might show that a large number of 2012 robberies occurred in the downtown precinct and that the average family income of persons who live in downtown Chicago was substantially higher than that in other precincts in the city. We might be tempted to conclude that high-income downtown residents are more likely to be robbed than are people who live in other parts of the city—that robbers select richer victims.

In reaching such a conclusion, we run the risk of committing the ecological fallacy by using observations aggregated into police precincts to make some statement about the socioeconomic status of individual robbery victims. We refer to this as a fallacy because lower-income people who did not live in the downtown area might also have been robbed there in 2012. Victims might be commuters to jobs in the Loop, people visiting downtown theaters or restaurants, passengers on subway or elevated train platforms, or homeless persons who are not counted by the census. Our problem is that we examined police precincts as our units of analysis, but we wish to draw conclusions about individual people.

The same problem will arise if we discover that incarceration rates are higher in states that have a large proportion of elderly residents. We will not know whether older people are actually imprisoned more often. Or, if we find higher suicide rates in cities with large nonwhite populations, we cannot be sure whether more nonwhites than whites committed suicide.

Don't let these warnings against the ecological fallacy lead you to commit what is called an "individualistic fallacy." Some students approaching criminal justice research for the first time have trouble reconciling general patterns of attitudes and actions with individual exceptions they know of. If, for example, you read a newspaper story about a Utah resident visiting New York who is murdered on a subway platform, the fact remains that most visitors to New York and most subway riders are not at risk of murder. Similarly, mass media stories and popular films about drug problems in U.S. cities frequently focus on drug use and dealing among African Americans. But that does not mean that most African Americans

are drug users or that drugs are not a problem among whites.

The individualistic fallacy can be especially troublesome for beginning students of criminal justice. Newspapers, local television news, and television police dramas often present unusual or highly dramatized versions of crime problems and criminal justice policy. These messages may distort the way many people initially approach research problems in criminal justice.

Reductionism

A second concept relating to units of analysis is reductionism. Basically, reductionism is an overly strict limitation on the kinds of concepts and variables to be considered as causes in explaining the broad range of human behavior represented by crime and criminal justice policy. Economists may tend to consider only economic variables (marginal value, expected utility); sociologists may consider only sociological variables (values, norms, roles); psychologists may consider only psychological variables (personality types, compulsive personality disorder). For example, why did violent crime by juveniles decline from 1995 through 2002, after an increase from 1985 through about 1994? Was it the result of changes in family structure? Fluctuation in economic opportunities for teenagers? Diminished fascination with images of violence, power, and flashy lifestyles among drug dealers in popular media? Social scientists from different disciplines tend to look at some explanations for crime problems and ignore the others. Explaining crime solely in terms of economic factors is economic reductionism; explaining crime solely in terms of psychological factors is psychological reductionism.

Reductionism of any type tends to suggest that particular units of analysis or variables are more relevant than others. If we consider the changing family structure as the cause of increased juvenile crime, our unit of analysis will be families. An economist, though, might use the 50 states as the units of analysis and compare juvenile crime rates and economic conditions. A psychologist might choose individual juveniles as the units of analysis to determine how watching violent films affects personality development.

Like the ecological fallacy, reductionism involves the use of inappropriate units of analysis. The appropriate unit of analysis for a given research question is not always clear and is often debated by social scientists, especially across disciplinary boundaries.

The box titled "Units of Analysis in the National Youth Gang Survey" offers several additional examples of using inappropriate units of analysis. It also illustrates that lack of clarity about units of analysis in criminal justice results in part from difficulties in directly measuring the concepts we want to study.

Units of Analysis in Review

The purpose of this section has been to specify what is sometimes a confusing topic, in part because criminal justice researchers use a variety of different units of analysis. Although individual people are often the units of analysis, that is not always the case. Many research questions can more appropriately be answered through the examination of other units of analysis.

The concept of unit of analysis may seem more complicated than it needs to be. Understanding the logic of units of analysis is more important than memorizing some list of the units. It is irrelevant what we call a given unit of analysis—a group, a formal organization, or a social artifact. It is essential, however, that we be able to identify what our unit of analysis is. We must decide whether we are studying assaults or assault victims, police departments or police officers, courtrooms or judges, and prisons or prison inmates. Without keeping this point in mind, we run the risk of making assertions about one unit of analysis based on the examination of another.

To test your grasp of the concept of units of analysis, here are some examples of real research topics. See if you can determine the unit of analysis in each. (The answers are given at the end of this chapter.)

1. "Taking into account preexisting traffic fatality trends and several other relevant factors, the implementation of the emergency cellular telephone program resulted in a substantial and permanent reduction in the monthly percentage of alcohol-related

CHAPTER 4 General Issues in Research Design 97

2. "Like the ecological fallacy, reductionism involves the use of inappropriate units of analysis. The appropriate unit of analysis for a given research question is not always clear and is often debated by social scientists, especially across disciplinary boundaries."

The box titled "Units of Analysis in the National Youth Gang Survey" offers several additional examples of using inappropriate units of analysis. It also illustrates that lack of clarity about units of analysis in criminal justice results in part from difficulties in directly measuring the concepts we want to study.
3. "The survey robbery rate was highest in Canada and the Netherlands, and lowest in Scotland.... In 1999 the survey robbery rate was lowest in the United States" (Farrington, Langan, and Tonry, 2004:xii).
4. "On average, probationers were 31 years old, African American, male, and convicted of drug or property offenses. Most lived with family, and although they were not married, many were in exclusive relationships (44 percent) and had children (47 percent)" (MacKenzie et al., 1999:433).
5. "Seventy-five percent ($n = 158$) of the cases were disposed at district courts, and 3 percent ($n = 6$) remained pending. One percent of the control and 4 percent of the experimental cases were referred to drug treatment court" (Taxman and Ellis, 1999:42).
6. "The department's eight Field Operations Command substations (encompassing 20 police districts and 100 patrol beats) comprised the study's cross-sectional units. From January 1986 through July 1989, a period of 43 months, data were collected monthly from each substation for a total of 344 cases" (Kessler, 1999:346).
7. For most violent crimes, American rates are among the highest—along with Australia, Canada, Spain, and France—but not the highest. Chances of being robbed, being assaulted, or being the victim of a stranger rape are higher in several other Western countries" (Tonry, 1999:421).
8. "Of all the flies opened for Arabs, 32 percent were crimes against the person, whereas of



UNITS OF ANALYSIS IN THE NATIONAL YOUTH GANG SURVEY

In 1997, the third annual National Youth Gang Survey was completed for the Federal Office of Juvenile Justice and Delinquency Prevention (OJJDP). This survey reflects keen interest in developing better information about the scope of youth gangs and their activities in different types of communities around the country. As important and useful as this effort is, the National Youth Gang Survey—especially reports of its results—illustrates how some ambiguities can emerge with respect to units of analysis.

A variety of attempts, often creative, are used to gather information from or about active offenders. Partly this is because it is difficult to systematically identify offenders for research. Studying youth gangs presents more than the usual share of problems with units of analysis. Are we interested in gangs (groups), gang members (individuals), or offenses (social act-fact) committed by gangs?

Following methods developed in earlier years, the 1997 National Youth Gang Survey was based on a sample of law enforcement agencies. The sample was designed to represent different types of communities:

- all the files opened for Jews, 30 percent were in that category. Proportionally, more files of property and drug offenses were opened for Jews than for Arabs. In addition, Arabs are over-represented in public order offenses" (Mesch and Fishman, 1999:184).
9. "Stories about women perpetrating crimes against children were more likely to appear on the first page of the newspaper, 7.5% versus 1%" (Grabe et al., 2006:149).
10. "Approximately half of the [burglars] (54 percent) admitted to 50 or more lifetime burglaries.... Included in this group are 44 offenders who had committed at least 100 such crimes. At the other extreme are 11 individuals who had participated in nine or fewer 'residential break-ins' (Wright and Decker, 1994:13).

rural areas, suburban counties, small cities, and large cities. Questionnaires were mailed to the police chief for municipalities and to the sheriff for counties (National Youth Gang Center, 1999:3). Questions asked respondents to report on gangs and gang activity in their jurisdiction—municipality for police departments, and unincorporated service area for sheriff's departments. Here are examples of the types of questions included in the survey:

1. How many youth gangs were active in your jurisdiction?
2. How many active youth gang members were in your jurisdiction?
3. In your jurisdiction, what percent of street sales of drugs were made by youth gang members? [followed by list: powder cocaine, crack cocaine, marijuana, heroin, methamphetamine, other]
4. Does your agency have the following? [list of special youth gang units]

Notice the different units of analysis embedded in these questions. Seven are stated or implied.

The Time Dimension

Because time order is a requirement for causal inferences, the time dimension of research requires careful planning.

We saw earlier in this chapter how the time sequence of events and situations is a critical element in determining causation. Time is also involved in the generalizability of research findings. Do the descriptions and explanations that result from a particular study accurately represent the situation of 10 years ago or 10 years from now, or do they represent only the present state of affairs? In general, observations may be made more or less at one time point, or they may be deliberately stretched over a longer period. Observations made at more than one time point can look forward or backward.

1. Gangs: item 1
2. Gang members: items 2, 3
3. Jurisdiction (city or part of county area): items 1, 2, 3
4. Street sales of drugs: item 3
5. Drug types: item 3
6. Agency: item 4
7. Special unit: item 4

Now, consider some quotes from a summary report on the 1997 survey (National Youth Gang Center, 1999). Which ones do or do not reasonably reflect the actual units of analysis from the survey?

- "Fifty-one percent of survey respondents indicated that they had active youth gangs in their jurisdictions in 1997" (page 7)
- "Thirty-eight percent of jurisdictions in the Northeast, and 26 percent of jurisdictions in the Middle Atlantic regions reported active youth gangs in 1997" (extracted from Table 3, page 10)
- "Results of the 1997 survey revealed that there were an estimated 30,533 youth gangs and 815,986 gang members active in the United States in 1997" (page 13)
- "The percentage of street sales of crack cocaine, heroin, and methamphetamine conducted by

Cross-Sectional Studies (Snodgrass)

Many criminal justice research projects are designed to study some phenomenon by taking a cross section of it at one time and analyzing that cross section carefully. Exploratory and descriptive studies are often **cross-sectional studies**. A single U.S. census, for instance, is a study aimed at describing the U.S. population at a given time. A single wave of the National Crime Victimization Survey (NCVS) is a descriptive cross-sectional study that estimates how many people have been victims of crime in a given time.

A cross-sectional exploratory study might be conducted by a police department in the form of a survey that examines what residents believe to be the sources of crime problems in their

youth gang members varied substantially by region.... Crack cocaine sales involving youth gang members were most prevalent in the Midwest (38 percent), heroin sales were most prevalent in the Northeast (15 percent), and methamphetamine sales were most prevalent in the West (21 percent)." (page 27)

■ "The majority (66 percent) of respondents indicated that they had some type of specialized unit to address the gang problem." (page 33)

The youth gang survey report includes a number of statements and tables that inaccurately describe units of analysis. You probably detected examples of this in some of the statements shown here. Other statements accurately reflect units of analysis measured in the survey.

If you read the 1997 survey report and keep in mind our discussion of units of analysis, you will find more misleading statements and tables. This will enhance your understanding of units of analysis.

Source: Information drawn from the National Youth Gang Center (1999).

neighborhood. In all likelihood, the study will ask about crime problems in a single time frame, with the findings used to help the department explore various methods of introducing community policing.

Cross-sectional studies for explanatory or evaluation purposes have an inherent problem. We have seen that inferring cause requires that a cause precede an effect in time, but cross-sectional studies produce observations made at only one time. For example, a survey might ask respondents whether their home has been burglarized and whether they have any special locks on their doors, hoping to explain whether special locks prevent

Cross-sectional study Data collected at a single time point.

burglary. Because the questions about burglary victimization and door locks are asked at only one time, it is not possible to determine whether burglary victims installed locks after a burglary or whether special locks were already in place but did not prevent the crime. Some of the ways we can deal with the difficult problem of determining time order will be discussed in the section on approximating longitudinal studies.

Longitudinal Studies

Research projects known as **longitudinal studies** are designed to permit observations over an extended period. An example is a researcher who interviews a group of subjects every 6 months for 5 years. Analysis of newspaper stories about crime or numbers of prison inmates over time are other examples. In the latter instances, it is irrelevant whether the researcher's observations are made over the course of the actual events under study or at one time—for example, examining a year's worth of newspapers in the library or 10 years of annual reports on correctional populations.

Three special types of longitudinal studies should be noted here: trend, cohort, and panel studies. **Trend studies** look at changes within some general population over time. An example is a comparison of the FBI's Uniform Crime Report (UCR) figures over time, showing an increase in reported crime from 1960 through 1993 and then a decline through 2008. Or a researcher might want to know whether changes in sentences for certain offenses were followed by increases in the number of people imprisoned in state institutions. In this case, a trend study might examine annual figures for prison population over time, comparing totals for the years before and after new sentencing laws took effect.

Cohort studies examine more specific populations (cohorts) as they change over time. Typically, a cohort is an age group, such as those people born during the 1980s, but it can also be based on some other time grouping. Cohorts are often defined as a group of people who enter or leave an institution at the same time, such as persons entering a drug

Longitudinal study Data collected at multiple time points.

treatment center during July, offenders released from custody in 2008, or high school seniors in March 2010.

In a classic cohort study, Marvin Wolfgang and associates (Wolfgang, Figlio, and Sellin, 1972) studied all males born in 1945 who lived in the city of Philadelphia from their 10th birthday through age 18 or older. The researchers examined records from police agencies and public schools to determine how many boys in the cohort had been charged with delinquency or arrested, how old they were when first arrested, and what differences there were in school performance between delinquents and nondelinquents.

Panel studies are similar to trend and cohort studies except that observations are made on the same set of people on two or more occasions. The NCVS is a good example of a descriptive panel study. A member of each household selected for inclusion in the survey is interviewed seven times at 6-month intervals. The NCVS serves many purposes, but it was developed initially to estimate how many people were victims of various types of crimes each year. It is designed as a panel study so that persons can be asked about crimes that occurred in the previous 6 months, and two waves of panel data are combined to estimate the nationwide frequency of victimization over a 1-year period. Panel studies are often used in evaluation research, in which the same persons are interviewed both before and after a new program is introduced.

Longitudinal studies tend to be expensive and difficult to conduct, and panel studies face a special problem: panel attrition. Some of the respondents studied in the first wave of a study may not participate in later waves. The danger is that those who drop out of the study may differ in some way compared to those who remain in the study, and may thereby distort the results of the study. Suppose we are interested in evaluating the success of a new drug treatment program by conducting weekly drug tests on a panel of participants for a period of 10 months. Regardless of how successful the program appears to be after 10 months, if a substantial number of people drop out of our study, we can expect that treatment was less effective in keeping them off drugs.

Approximating Longitudinal Studies

It may be possible to draw conclusions about processes that take place over time even when only cross-sectional data are available. It is worth noting some of the ways to do that.

Sometimes, cross-sectional data imply processes that occur over time on the basis of simple logic. Consider, for example, a study of gun ownership and violence by Swiss researcher Martin Killias (1993). Killias compared rates of gun ownership as reported in an international crime survey to rates of homicide and suicide committed with guns. He was interested in the possible effects of gun availability on violence: Do nations with higher rates of gun ownership also have higher rates of gun violence?

Killias reasoned that inferring causation from a cross-sectional comparison of gun ownership and homicides committed with guns would be ambiguous. Gun homicide rates could be high in countries with high gun ownership rates because the availability of guns was higher. Or people in countries with high gun ownership rates could have bought guns to protect themselves, in response to rates of homicide. Cross-sectional analysis would not make it possible to sort out the time order of gun ownership and gun homicides.

But does that reasoning hold for gun suicides? Killias argued that the time order in a relationship between gun ownership and gun suicides is less ambiguous. It makes much more sense that suicides involving guns are at least partly a result of gun availability. But it is not reasonable to assume that people might buy guns in response to high gun suicide rates.

Logical inferences like that made by Killias may also be sometimes made whenever the time order of variables is clear. For example, if we discover in a cross-sectional study of high school students that men are more likely than women to smoke marijuana, we can conclude that gender might affect the propensity to use marijuana, not the other way around. Thus, even though our observations are made at only one time, we are justified in drawing conclusions about processes that take place across time.

Retrospective and Prospective Studies

Retrospective research, which asks people to recall their pasts, is another common way of approximating observations over time. In a study of recidivism, for example, we might select a group of prison inmates and analyze their history of delinquency or crime. Or suppose we are interested in whether college students convicted of drunk driving are more likely to have parents with drinking problems than college students with no drunk-driving record. Such a study is retrospective because it focuses on the histories of college students who have or have not been convicted of drunk driving.

The danger in this technique is evident. Sometimes, people have faulty memories. Retrospective recall is one way of approximating observations across time, but it must be used with caution. Retrospective studies that analyze records of past arrests or convictions suffer from different problems: records may be unavailable, incomplete, or inaccurate.

A more fundamental problem in retrospective research hinges on how subjects are selected and how subject selection affects the kinds of questions such studies can address.

Imagine that you are a juvenile court judge, and you're troubled by what appears to be a large number of child abuse cases in your court. Talking with a juvenile caseworker, you wonder whether the parents of these children were abused or neglected during their own childhood. Together, you formulate a hypothesis about the intergenerational transmission of violence: victims of childhood abuse later abuse their own children. How might you go about investigating that hypothesis?

Given your position as a judge who regularly sees abuse victims, you will probably consider a retrospective approach that examines the backgrounds of families appearing in your court. Let's say you and the caseworker plan to investigate the family backgrounds of 20 abuse victims who appear in your court during the next 3 months. The caseworker consults with a clinical psychologist from the local university and obtains copies of a questionnaire, or "protocol," that has been used by researchers to study the families of child abuse victims. After interviewing the families of 20

victim, the caseworker reports to you that 18 of the 20 child victims have a mother or father who was abused as a child. It seems safe to conclude that your hypothesis about the intergenerational transmission of violence is strongly supported, because 90 percent (18 out of 20) of abuse/neglect victims brought before your court come from families with a history of child abuse.

Think for a moment about how you approached the question of whether child abuse breeds child abuse. You began with abuse victims and retrospectively established that many of their parents had been abused. However, this is different from the question of how many victims of childhood abuse later abuse their own children. That question requires a **prospective** approach, in which you begin with childhood victims and then determine how many of them later abuse their own children.

To clarify this point, let's shift from the hypothetical study to actual research that illustrates the difference between prospective and retrospective approaches to the same question. Rosemary Hunter and Nancy Kilstrom (1979) conducted a study of 255 infants and their parents. The researchers began by selecting families of premature infants in a newborn intensive care unit. Interviews with the parents of 255 infants revealed that either the mother or the father in 49 of the families had been the victim of abuse or neglect; 206 families revealed no history of abuse. In a prospective follow-up study, Hunter and Kilstrom found that within 1 year, 10 of the 255 infants had been abused. Nine of those 10 infant victims were from the 49 families with a history of abuse, while 1 abused infant was from the 206 families with no background of abuse.

Figure 4.2A illustrates these prospective results graphically. Infants in 18 percent (9 out of 49) of families with a history of abuse showed signs of abuse within 1 year of birth, while less than 1 percent of infants born to parents with no history of abuse were themselves abused within 1 year. Although that is a sizable difference, notice that the 18 percent figure for continuity of abuse is very similar to the 19 percent rate of abuse discovered in the histories of all 255 families.

Now consider what Hunter and Kilstrom would have found if they had begun with the 10 abused

infants at time 2 and then checked their family back grounds. Figure 4.2B illustrates this retrospective approach. A large majority of the 10 infant victims (90 percent) had parents with a history of abuse.

You probably realize by now that the prospective and retrospective approaches address fundamentally different questions, even though the questions may appear similar on the surface:

Prospective: What percentage of abuse victims later abuse their children? (18 percent; Figure 4.2A)

Retrospective: What percentage of abuse victims have parents who were abused? (90 percent; Figure 4.2B)

In a study of how child abuse and neglect affect drug use, Cathy Spatz Widom and associates (Widom, Weiler, and Cotler, 1999) present a similar contrast of prospective and retrospective analysis. Looking backward, 75 percent of subjects with a drug abuse diagnosis in semiclinical interviews were victims of childhood abuse or neglect. Looking forward, 35 percent of childhood victims and 34 percent of nonvictims had a drug abuse diagnosis.

More generally, Robert Sampson and John Laub (1993:14) comment on how retrospective and prospective views yield different interpretations about patterns of criminal offending over time:

Looking back over the careers of adult criminals exaggerates the prevalence of stability. Looking forward from youth reveals the success and failures, including adolescent delinquents who go on to be normal functioning adults. This is the paradox noted [by Lee Robins] earlier: adult criminality seems to be always preceded by childhood misconduct, but most conduct-disordered children do not become antisocial or criminal adults (Robins, 1978). (emphasis in original)

Our intention here is not to suggest that retrospective studies have no value. Rather, we want to point out how the time dimension is linked to the framing of research questions. A retrospective approach is limited in its ability to reveal how causal processes unfold over time. A retrospective approach is therefore not well suited to answer questions such as how many childhood victims of abuse or neglect later abuse their own children.

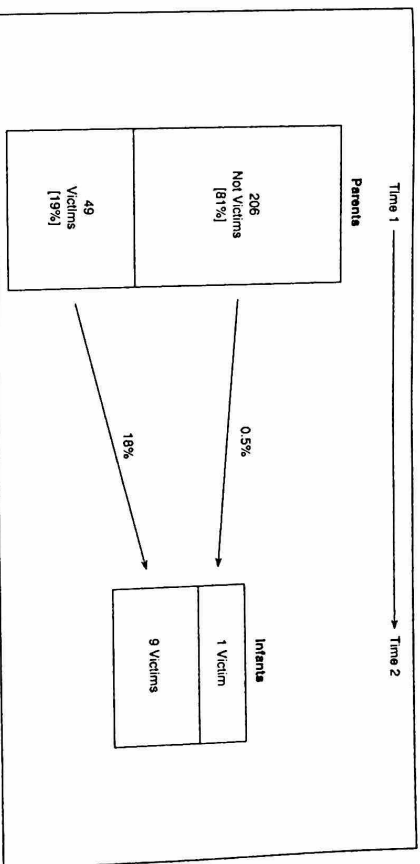


Figure 4.2A Prospective Approach to a Subject

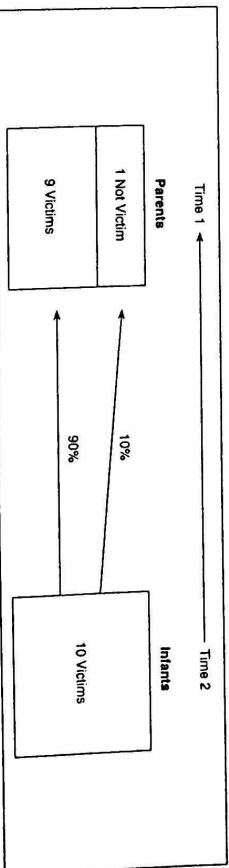


Figure 4.2B Retrospective Approach to a Subject

Source: Adapted from Hunter and Kilstrom (1979), as suggested by Widom (1989).

A retrospective study can be used, however, to compare whether childhood victims are more likely than nonvictims to have a history of abuse in their family background.

The Time Dimension Summarized

Joel Devine and James Wright (1993:19) offer a clever metaphor that distinguishes longitudinal studies from cross-sectional ones. Think of a cross-sectional study as a snapshot, a trend study as a slide show, and a panel study as a motion

picture. A cross-sectional study, like a snapshot, produces an image at one point in time. This can provide useful information about crime—burglary, for example—at a single time, perhaps in a single place. A trend study is akin to a slide show—a series of snapshots, in sequence over time. By viewing a slide show, we can tell how some indicator—change in burglary rates—varies over time. But a trend study is usually based on aggregate information. It can tell us something about aggregations of burglary over time, but not, for instance, whether the same people are committing burglaries at an increasing or decreasing

rate, or whether there are more or fewer burglars with a relatively constant rate of crime commission. A panel study, like a motion picture, can capture moving images of the same individuals and give us information about individual rates of offending over time.



CAUSE, UNITS, AND TIME

Causation in Traffic Stops

What causes observed proportions of traffic stops by race of driver? That is the central question confronted by researchers and public officials who tried to sort out the problem of racial profiling. Studies in three states framed these questions somewhat differently, but each was concerned with one descriptive question, then a family of causal questions:

1. Are there racial or ethnic disparities in the traffic stops?
2. If yes, what causes these disparities?
 - Over discrimination by police?
 - Something else?

The first question refers to covariation as a criterion for cause. It is necessary to show disparities, or disproportionality by race before determining that discrimination exists. In other words, if minority motorists are not being stopped or ticketed in greater proportion to their numbers in the population, then discrimination does not exist.

Regarding the second question, most news reports and horror stories assumed some form of discrimination was responsible for any observed disparities. In fact, some researchers seemed to believe that discrimination was the only possible reason for disparities in traffic stops (for example, Lamberth, 2003). Others recognized that different explanations could be at work. For example, Smith et al. wrote:

These, then, are some of the ways time figures in criminal justice research and some of the ways researchers have learned to cope with it. In designing any study, you need to look at both the explicit and implicit assumptions they are making about time. Are you interested in describing some

You should recognize this statement as an implicit challenge to the internal validity of inferring that discrimination causes disparity. In fact, Smith and associates found a number of other variables to be associated with disparities in traffic stops.

Maxfield and Kelling (2005:5–6) addressed a more narrow research question: Why was the proportion of minority drivers stopped substantially higher for one segment of the New Jersey Turnpike than for other segments? Their research considered five possible causal factors in addition to actual discrimination:

- **Deployment:** Do more troopers happen to be deployed in areas with high proportions of minority drivers?
- **Behavior:** In comparison to whites, do minority motorists disproportionately engage in illegal driving practices that bring them to the attention of the New Jersey State Police?
- **Instrumentation:** Are New Jersey State Police troopers now more accurately and/or more consistently reporting the race and ethnicity of traffic violators in comparison to past practices?
- **Evasion:** Are white traffic offenders on the turnpike more able to avoid detection by the New Jersey State Police than minority traffic offenders?
- **Interaction:** Is the continuity over time of racially disproportionate stops in the stop data caused by some mix of the above hypotheses?

Similar alternative explanations were considered by most other researchers in most other areas of the country. In later chapters we describe steps taken to address these questions.

process that occurs over time, such as whether mandatory jail sentences reduce drunk driving? Or are you simply going to describe how many people were arrested for drunk driving in the past year? If you want to describe a process that occurs over time, will you be able to make observations

Units

Studying traffic enforcement and the possibility of race-based discrimination involves several different units of analysis. The studies done in Pennsylvania and North Carolina collected data on several units, moving from the least to the greatest level of aggregation:

- traffic stops: driver characteristics and reasons for stop
- trooper: aggregate characteristics of all drivers and stops; plus data on individual trooper
- small-area segment: aggregate characteristics of all troopers, drivers, and stops; plus data on road segment and individual area
- station or county: aggregates from lower-level units; plus summary data on county residents

Then, of course, studies done in different states reflected the laws and other features of each state. For example, in Pennsylvania troopers are prohibited from stopping cars traveling less than 7 mph above the posted speed limit. New Jersey has no stated policy, but informal norms tolerate speeding up to about 15 mph over the posted limit on many expressways.

CHAPTER 4 General Issues in Research Design 105

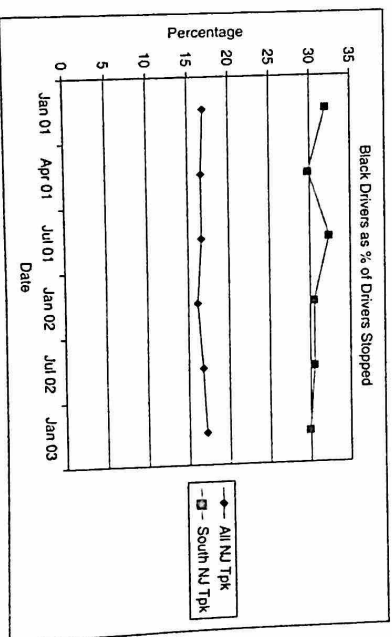
at different points in the process, or will you have to approximate such observations, drawing logical inferences from what you can observe now?

Our “Putting It All Together” example for this chapter illustrates many of the general issues in research design we have discussed so far.

Time Dimension

The time dimension revealed certain trends in traffic enforcement. In North Carolina, researchers found a sharp decline in traffic enforcement during 1999, when the state legislature was considering laws to monitor the possibility of racial profiling (Smith et al. 2003:28–29). This documents what Engel and associates term “disengagement,” where police cut back on enforcement activities to avoid getting into trouble.

The time dimension was one factor triggering research in New Jersey. The figure below shows, over a 2-year period, what proportion of drivers stopped for moving violations was black. Two generally stable trend lines are presented. The lower one shows data for the entire New Jersey Turnpike, while the upper line represents the southern segment, about 50 miles long. You can see that the proportion of stopped drivers who were black is about twice as high for the southern segment as for the entire turnpike. The size and stability of this difference was one of the main reasons for launching the New Jersey study. The time dimension here played a role in documenting the persistence of race disparities over time.



© Cengage Learning. All rights reserved.

Answers to the Units-of-Analysis Exercise

1. Social artifacts (alcohol-related fatal crashes)
2. Groups (cities and counties)
3. Groups (countries)
4. Individuals (probationers)
5. Social artifacts (court cases)
6. Organizations (police substations)
7. Groups (countries)
8. Social artifacts (offense files)
9. Social artifacts (newspaper stories)
10. Individuals (burglars)

SUMMARY

- Explanatory scientific research centers on the notion of cause and effect.
- Most explanatory social research uses a probabilistic model of causation. X may be said to cause Y if it is seen to have some influence on Y .
- X is a necessary cause of Y if Y cannot happen without X having happened. X is a sufficient cause of Y if Y always happens when X happens.
- Three basic requirements determine a causal relationship in scientific research: (1) The independent variable must occur before the dependent variable, (2) the independent and dependent variables must be empirically related to each other, and (3) the observed relationship cannot be explained away as the effect of another variable.
- When scientists consider whether causal statements are true or false, they are concerned with the validity of causal inference.
- Four classes of threats to validity correspond to the types of questions researchers ask in trying to establish cause and effect. Threats to statistical conclusion validity and internal validity arise from bias. External and construct validity threats may limit our ability to generalize from an observed relationship.
- A scientific realist approach to examining mechanisms in context brings ethnographic and nomothetic approaches to causation.
- Units of analysis are the people or things whose characteristics researchers observe, describe, and explain. The unit of analysis in criminal justice research is often the individual person, but it may also be a group, organization, or social artifact.

- Researchers sometimes confuse units of analysis resulting in the ecological fallacy or the individualistic fallacy.
- Cross-sectional studies are those based on observations made at one time. Although such studies are limited by this characteristic, inferences can often be made about processes that occur over time.
- Longitudinal studies are those in which observations are made at many times. Such observations may be made of samples drawn from general populations (trend studies), samples drawn from more specific subpopulations (cohort studies), or the same sample of people each time (panel studies).
- Retrospective studies can sometimes approximate longitudinal studies, but retrospective approaches must be used with care.

KEY TERMS

- Cohort study (p. 100)
- Construct validity (p. 89)
- Cross-sectional study (p. 99)
- Ecological fallacy (p. 95)
- External validity (p. 89)
- Internal validity (p. 88)
- Longitudinal study (p. 100)
- Panel study (p. 100)
- Probabilistic (p. 85)
- Prospective (p. 102)
- Retrospective (p. 101)
- Scientific realism (p. 92)
- Statistical conclusion validity (p. 87)
- Trend study (p. 100)
- Units of analysis (p. 92)
- Validity (p. 87)
- Validity threats (p. 87)

REVIEW QUESTIONS AND EXERCISES

1. Discuss one of the following statements in terms of what you have learned about the criteria of causation and threats to the validity of causal inference. What cause-and-effect relationships are implied? What are some alternative explanations?
 - a. Guns don't kill people; people kill people.
 - b. Capital punishment prevents murder.
 - c. Marijuana is a gateway drug that leads to the use of other drugs.

2. Several times, we have discussed the relationship between drug use and crime. Describe the conditions that would lead us to conclude that drug use is:
 - a. A necessary cause
 - b. A sufficient cause
 - c. A necessary and sufficient cause
3. In describing different approaches to the time dimension, criminologist Lawrence Sherman (1995) claims that cross-sectional studies can show differences and that longitudinal studies can show change. How does this statement relate to the three criteria for inferring causation?

CHAPTER 4 General Issues in Research Design 107

4. William Julius Wilson (1996:167) cites the following example of why it's important to think carefully about units and time. Imagine a 13-bed hospital, in which 12 beds are occupied by the same person for 1 year. The other hospital bed is occupied by 52 people, each staying 1 week. At any given time, 92 percent of beds are occupied by long-term patients (12 out of 13), but over the entire year, 81 percent of patients are short-term patients (52 out of 64). Discuss the implications of a similar example, in which "jail cell" is substituted for "hospital bed."

CHAPTER 7

Experimental and Quasi-Experimental Designs

You'll learn about the experimental approach to social science research. We'll consider a wide variety of experimental and other designs available to criminal justice researchers.



Learning Objectives

1. Recognize that experiments are well suited for the controlled testing of causal processes and for some evaluation studies.
2. Describe how the classical experiment tests the effect of an experimental stimulus on some dependent variable through the pretesting and posttesting of experimental and control groups.
3. Understand that a group of experimental subjects need not be representative of some larger population but that experimental and control groups must be similar to each other.
4. Describe how random assignment is the best way to achieve comparability in the experimental and control groups.
5. Describe how the classical experiment with random assignment of subjects guards against most of the threats to internal validity.
6. Understand that the controlled conditions under which experiments take place may restrict our ability to generalize results to real-world constructs or to other settings.
7. Recognize how the classical experiment may be modified by changing the number of experimental and control groups, the number and types of experimental stimuli, and the number of pretest or posttest measurements.
8. Know the reasons that quasi-experiments are conducted when it is not possible or desirable to use an experimental design, and be able to describe different categories of quasi-experiments.
9. Understand the differences between case-oriented and variable-oriented research. Time-series

designs and case studies are examples of variable-oriented research, in which a large number of variables are examined for one or a few cases.

CHAPTER 7 Experimental and Quasi-Experimental Designs 167

10. Be able to describe how experiments and quasi-experiments can be customized by using design building blocks to suit particular research purposes.

Introduction 168

The Classical Experiment 169

- Independent and Dependent Variables 169
- Pretesting and Posttesting 170
- Experimental and Control Groups 170
- Double-Blind Experiments 172
- Selecting Subjects 173
- Random Assignment 173

Experiments and Causal Inference 174

- Experiments and Threats to Validity 175
- Threats to Internal Validity 175
- Ruling Out Threats to Internal Validity 178
- Generalizability and Threats to Validity 179
- Threats to Construct Validity 179
- Threats to External Validity 181
- Threats to Statistical Conclusion Validity 181

Variations in the Classical Experimental Design 182

Quasi-Experimental Designs 183

- Nonequivalent-Groups Designs 184
- Cohort Designs 187
- Time-Series Designs 188
- Variations in Time-Series Designs 191
- Variable-Oriented Research and Scientific Realism 192

PUTTING IT ALL TOGETHER: EVALUATING EXPLANATIONS FOR DISPROPORTIONALITY IN TRAFFIC STOPS 194

Experimental and Quasi-Experimental Designs Summarized 196

Summary 196

An Experiment in Spear Phishing

We're willing to bet that any person reading this book has received many offers of riches to be bestowed by bankers and oil ministers from England, Nigeria, Ukraine, and just about everywhere else. Maybe you've wondered how anyone could be so glibble to be taken in by this form of "phishing." Or maybe you, like many people, have responded to or wondered about more sophisticated email phishing attempts. Michael Maxfield was temporarily taken in by the following message: "Your new computer has shipped and will arrive in two days! Click here to review or change payment and delivery information."

Computer scientists at Indiana University conducted an experiment to learn something about what

influences people to respond to phishing. Their basic research question was "How easily and effectively can a phisher exploit social networking data found on the Internet to increase the yield of a phishing attack?" (Jagatic et al., 2007).

Tom Jagatic and colleagues searched for publicly available information about 23,000 Indiana students, seeking students who participated in social networking sites such as Facebook. Approximately 1,700 students met the researchers' criteria, which included having several friends or circles of friends on the sites.

They then sent two versions of phishing emails to different groups of students. One group received a

treatment condition message that mimicked someone in the student's circle of friends; this is sometimes known as "spear phishing." The other group received a control message that contained no information, suggesting the sender was part of the receiver's network. For example, assume Maxfield was selected as a treatment subject, and Earl Bobbie was among his circle of friends on Facebook. The researchers might have sent Maxfield a message that appeared to be from Bobbie, something like: "Hi Mikel! I just found this great new website that has copies of hundreds of survey questionnaires. Check it out at: megasurveys.com. Take care, Earl." Or, if Maxfield was a control subject in this experiment, a message might have read: "Dear Professor: Please click on the link below to learn more about our collection of survey questionnaires." In either case, the message would have redirected Maxfield to some other site—harmless, but clearly not a compilation of survey questionnaires. Which message—treatment or control—do you think would be more likely to prompt recipients to click on the enclosed link?

Jagatic and associates found that a whopping 72 percent of people receiving the treatment message followed the link, compared to 16 percent of those who got the control message. In addition, spear phishing was more effective in cross-gender messages. Females more often clicked on links they believed were from their male friends than to links from female friends; male recipients were similarly duped when they thought messages came from female friends.

This is an excellent example of an experiment. It involved an unambiguous intervention that could be easily controlled: applied to a treatment group, but not to a control group. Researchers could readily determine which subjects were in each group and assign them in a random, unbiased way. As you read this chapter, compare this example to our general discussion of designs and to other examples. What features of the setting for this study facilitated an experimental approach? How is the setting different from the environment in which criminal justice agencies routinely work?

Introduction

Experimentation is an approach to research best suited for explanation and evaluation.

Research design in the most general sense involves devising a strategy for finding out something. We'll first discuss the experiment as a mode of scientific observation in criminal justice research. In his classic book, *The Conduct of Inquiry*, Abraham Kaplan (1964:144) described experimentation as "a process of observation, to be carried out in a situation expressly brought about for that purpose." At base, experiments involve (1) taking action and (2) observing the consequences of that action. Social science researchers typically select a group of subjects, do something to them, and observe the effect of what was done.

It is worth noting at the outset that experiments are often used in nonscientific human inquiry as well. We experiment copiously in our attempts to develop a more generalized understanding about the world we live in. We learn many skills through experimentation: riding a bicycle, driving a car, swimming, and so forth. Students discover how much studying is required for academic success through experimentation. Professors learn how much preparation is required for successful lectures through experimentation.

Returning to social science applications, experiments are especially well suited to research projects that involve relatively well-defined concepts and propositions. A further requirement is the ability to control the conditions under which research is conducted. The traditional model of science, discussed in Chapter 2, and the experimental model are closely related.

Experimentation is best suited for hypothesis testing in explanatory or evaluation studies. Suppose we are interested in studying alcohol abuse among college students and in discovering ways to reduce it. We might hypothesize that acquiring an understanding about the health consequences of binge drinking and long-term alcohol use will have the effect of reducing alcohol abuse. We can test this hypothesis experimentally. To begin, we

might ask a group of experimental subjects how much beer, wine, or spirits they drank on the previous day and how frequently, in an average week, they consume alcohol for the specific purpose of getting drunk. Next, we might show these subjects a video depicting the various physiological effects of chronic drinking and binge drinking. Finally—say, 1 month later—we might again ask the subjects about their use of alcohol in the previous week to determine whether watching the video actually reduced alcohol use.

Because experiments are best suited for hypothesis testing, they can also be appropriate in the study of criminal justice policy. In Chapter 2, we discussed the logical similarity between hypotheses and criminal justice policies, noting that evaluation research is conceptually equivalent to hypothesis testing. The experimental model, therefore, can be a useful design for evaluating criminal justice policy.

You might typically think of experiments as being conducted in laboratories under carefully controlled conditions. Although this is often true in the natural sciences, few social scientific experiments take place in laboratory settings. The most notable exception to this occurs in the discipline of psychology, in which laboratory experiments are common. Criminal justice experiments are almost always conducted in field settings outside the laboratory.

The Classical Experiment

Variables, time order, measures, and groups are the central features of the classical experiment.

Like much of the vocabulary of research, the word *experiment* has acquired both a general and a specialized meaning. So far, we have referred to the general meaning, defined by David Farrington, Lloyd Ohlin, and James Q. Wilson (1986:65) as "a systematic attempt to test a causal hypothesis about the effect of variations in one factor (the independent variable) on another (the dependent variable).... The defining feature of

an experiment lies in the control of the independent variable by the experimenter." In a narrower sense, the term *experiment* refers to a specific way of structuring research, usually called the **classical experiment**. In this section, we examine the requirements and components of the classical experiment. Later in the chapter, we will consider designs that can be used when some of the requirements for classical experiments cannot be met.

The most conventional type of experiment, in the natural and the social sciences, involves three major pairs of components: (1) independent and dependent variables, (2) pretesting and posttesting, and (3) experimental and control groups. We will now consider each of those components and the way they are put together in the execution of an experiment.

Independent and Dependent Variables

Essentially, an experiment examines the effect of an **independent variable** on a **dependent variable**. Typically, the independent variable takes the form of an experimental stimulus that is either present or absent—that is, a dichotomous variable, having two attributes. That need not be the case, however, as subsequent sections of this chapter will show. In the example concerning alcohol abuse, how often subjects used alcohol is the dependent variable and exposure to a video about alcohol's effects is the independent variable. The researcher's hypothesis suggests that levels of alcohol use depend, in part, on understanding its physiological and health effects. The purpose of the experiment is to test the validity of this hypothesis. In the opening vignette, "An Experiment in Spear Phishing," what kind of email message a person receives is the independent variable, and whether or not subjects click on a link is the dependent variable.

Classical experiment Research design with three components: pre- and posttests; experimental and control groups; and random assignment to groups.

Keep in mind, however, that a given variable might serve as an independent variable in one experiment and as a dependent variable in another. For example, alcohol abuse is the dependent variable in our example, but it might be the independent variable in an experiment that examines the effects of alcohol abuse on academic performance.

In the terms of our discussion of cause and effect in Chapter 4, the independent variable is the cause and the dependent variable is the effect. Thus, we might say that watching the video *causes* a change in alcohol use or that reduced alcohol use is an *effect* of watching the video. A socially targeted email message causes people to click on an embedded link, and clicking on the link is the effect of the email message.

It is essential that both independent and dependent variables be operationally defined for the purposes of experimentation. Such operational definitions might involve a variety of observation methods. Responses to a questionnaire, for example, might be the basis for defining self-reported alcohol use on the previous day. Alternatively, alcohol use by subjects could be measured with Breathalyzer or blood alcohol tests.

Conventionally, in the experimental model, the dependent and independent variables are operationally defined before the experiment begins. However, as we will see in connection with survey research and other methods, it is sometimes appropriate to first make a variety of observations during data collection and then determine the most useful operational definitions of variables during later analysis. Ultimately, however, experimentation requires specific and standardized measurements and observations.

Prestesting and Posttesting

In the simplest experimental design, subjects are measured on a dependent variable (pretested), exposed to a stimulus that represents an

Control group Subjects in an experiment who do not receive the experimental treatment.

independent variable, and then remeasured on the dependent variable (posttested). Differences noted between the first and second measurements on the dependent variable are then attributed to the influence of the independent variable.

In our example of alcohol use, we might begin by prestesting the extent of alcohol use among our experimental subjects. Using a questionnaire, we measure the extent of alcohol use reported by each individual and the average level of alcohol use for the whole group. After showing subjects the video on the effects of alcohol, we administer the same questionnaire again. Responses given in this posttest permit us to measure the subsequent extent of alcohol use by each subject and the average level of alcohol use of the group as a whole. If we discover a lower level of alcohol use on the second administration of the questionnaire, we might conclude that the video indeed reduced the use of alcohol among the subjects.

In the experimental examination of behaviors such as alcohol use, we face a special practical problem relating to validity. As you can imagine, the subjects might respond differently to the questionnaire the second time, even if their level of drinking remained unchanged. During the first administration of the questionnaire, the subjects may have been unaware of its purpose. By the time of the second measurement, however, they may have figured out the purpose of the experiment, become sensitized to the questions about drinking, and changed their answers. Thus, the video might *seem* to have reduced alcohol abuse, although, in fact, it did not.

This is an example of a more general problem that plagues many forms of criminal justice research: the very act of studying something may change it. Techniques for dealing with this problem in the context of experimentation are covered throughout the chapter.

Experimental and Control Groups

The traditional way to offset the effects of the experiment itself is to use a **control group**. Social scientific experiments seldom involve only the

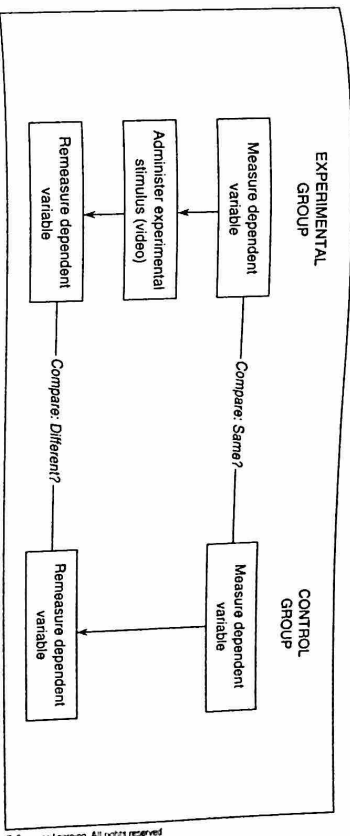


Figure 7.1 Basic Experimental Design

observation of an **experimental group** to which a stimulus has been administered. Researchers also observe a control group to which the experimental stimulus has not been administered.

In our example of alcohol abuse, two groups of subjects are examined. To begin, each group is administered a questionnaire designed to measure their alcohol use in general and binge drinking in particular. Then only one of the groups—the experimental group—is shown the video. Later, the researcher administers a posttest of alcohol use to both groups. Figure 7.1 illustrates this basic experimental design.

Using a control group allows the researcher to control for the effects of the experiment itself. If participation in the experiment leads the subjects to report less alcohol use, that should occur in both the experimental and the control groups. If, on the one hand, the overall level of drinking exhibited by the control group decreases between the pretest and posttest as much as for the experimental group, then the apparent reduction in alcohol use must be a function of the experiment or of some external factor, not a function of watching the video specifically. In this situation, we can conclude that the video did not cause any change in alcohol use.

If, on the other hand, drinking decreases only in the experimental group, then we can be more

confident in saying that the reduction is a consequence of exposure to the video (because that's the only difference between the two groups). Or, alternatively, if drinking decreases more in the experimental group than in the control group, then that, too, is grounds for assuming that watching the video reduced alcohol use.

The need for control groups in social scientific research became clear in a series of employee satisfaction studies conducted by Fritz J. Roethlisberger and William J. Dickson (1939) in the late 1920s and early 1930s. They studied working conditions in the telephone "bank writing room" of the Western Electric Works in Chicago, attempting to discover what changes in working conditions would improve employee satisfaction and productivity.

To the researchers' great satisfaction, they discovered that making working conditions better consistently increased satisfaction and productivity. When the workroom was brightened by better lighting, for example, productivity went up. Lighting was further improved, and productivity went up again. To substantiate their scientific conclusion, the researchers then dimmed the lights. *Productivity again improved!*

Experimental group Subjects in an experiment who receive the experimental treatment.

It became evident then that the wiring room workers were responding more to the attention given them by the researchers than to improved working conditions. As a result of this phenomenon, often called the "Hawthorne effect," social researchers have become more sensitive to and cautious about the possible effects of experiments themselves. The use of a proper control group—studied intensively, with the working conditions otherwise unchanged—would have revealed this effect in the wiring room study.

The need for control groups in experimentation has been most obvious in medical research. Time and again, patients who participated in medical experiments appeared to improve, but it was unclear how much of the improvement came from the experimental treatment and how much from the experiment. Now, in testing the effects of new drugs, medical researchers frequently administer a placebo (for example, sugar pills) to a control group. Thus, the control-group patients believe that they, like members of the experimental group, are receiving an experimental drug—and they often improve! If the new drug is effective, however, those who receive that drug will improve more than those who received the placebo.

In criminal justice experiments, control groups are important as a guard against the effects not only of the experiments themselves but also of events that may occur outside the laboratory during the course of experiments. Suppose the alcohol use experiment was being conducted on your campus, and at that time, a popular athlete was hospitalized for acute alcohol poisoning after he and a chum drank a bottle of rum. This event might shock the experimental subjects and thereby decrease their reported drinking. Because such an effect should happen about equally for members of the control and experimental groups, lower levels of reported alcohol use in the experimental group than in the control group would again demonstrate the impact of the experimental stimulus: watching the video that describes the health effects of alcohol abuse.

Sometimes, an experimental design requires more than one experimental or control group. In the case of the alcohol video, for example, we

might also want to examine the impact of participating in group discussions about why college students drink alcohol, with the intent of demonstrating that peer pressure may promote drinking by people who would otherwise abstain. We might design our experiment around three groups. One group would see the video and participate in the group discussions; another would only see the video, and still another would only participate in group discussions; the control group would do neither. With this kind of design, we could determine the impact of each stimulus separately, as well as their combined effect.

Double-Blind Experiments

As we saw with respect to medical experimentation, patients sometimes improve when they think they are receiving a new drug; thus, it is often necessary to administer a placebo to a control group.

Sometimes, experimenters have this same tendency to prejudge results. In medical research, the experimenters may be more likely to "observe" improvements among patients who receive the experimental drug than among those receiving the placebo. That would be most likely, perhaps, for the researcher who developed the drug. A double-blind experiment eliminates this possibility because neither the subjects nor the experimenters know which is the experimental group and which is the control. In medical experiments, those researchers who are responsible for administering the drug and for noting improvements are not told which subjects receive the drug and which receive the placebo. Thus, both researchers and subjects are "blind" with respect to who is receiving the experimental drug and who is getting the placebo. Another researcher knows which subjects are in which group, but that person is not responsible for administering the experiment.

An example of a double-blind experiment is a study of domestic violence policing units in Broward County Florida and Portland, Oregon (Feder, Jolin, and Feyerherm, 2000; Jolin et al., 1998). The researchers evaluated an intervention designed to reduce repeat offenses against victims of domestic violence by instructing them

in strategies for "keeping safe" and putting more resources into the investigation of domestic violence incidents.

In the experimental evaluation, after misdemeanor offenses for which male offenders had been arrested and taken to jail, their female victims were randomly assigned to a treatment group that received supplementary services or to a control group that did not receive the intervention. Cases in the treatment group were more intensively investigated, and more thorough efforts were made to collect evidence. Victims in the treatment group received additional services, including assistance in developing safety plans, information on how to access a variety of services, and transportation to service providers when necessary. Staff randomly assigned cases to the treatment or control groups, but assignment was coded so the staff were not aware of which victims were to receive supplementary services and which were not. Likewise, victims were not informed whether they were in the treatment or the control group. This double-blind experiment was, therefore, able to test the effects of supplementary domestic violence services without being contaminated by the expectations of victims or researchers.

In social scientific experiments, as in medical experiments, the danger of experimenter bias is further reduced to the extent that the operational definitions of the dependent variables are clear and precise. Thus, in the domestic violence experiment, researchers were less likely to unconsciously bias their interpretation of reports of repeat domestic violence than they were to unconsciously bias their subjective assessment of how satisfied victims were with services provided to them.

Selecting Subjects

Before beginning an experiment, we must make two basic decisions about who will participate. First, we must decide on the target population—the group to which the results of our experiment will apply. If our experiment is designed to determine, for example, whether restitution is more effective than probation in reducing recidivism,

our target population is some group of persons convicted of crimes. In our hypothetical experiment about the effects of watching a video on the health consequences of alcohol abuse, the target population might be college students.

Second, we must decide how particular members of the target population will be selected for the experiment. Ideally, the methods used to select subjects must meet the scientific norm of generalizability; it should be possible to generalize from the sample of subjects actually studied to the population those subjects represent.

The cardinal rule of subject selection for experimentation is the comparability of the experimental and control groups. Ideally, the control group represents what the experimental group would have been like if it had not been exposed to the experimental stimulus. It is important, therefore, that the experimental and control groups be as similar as possible.

Random Assignment

Having recruited a group of subjects, under the classic experiment we would randomly assign those subjects to either the experimental or the control group. This might be accomplished by numbering all the subjects serially and selecting numbers by means of a random-number table. Or we might assign the odd-numbered subjects to the experimental group and the even-numbered subjects to the control group.

Random assignment is a central feature of the classical experiment. The most important characteristic of random assignment (sometimes referred to as randomization) is that it produces experimental and control groups that are statistically equivalent. Put another way, random assignment reduces sources of systematic bias in assigning subjects to groups. The basic principle is simple: if subjects are assigned to experimental and control groups through a random process such as flipping a coin, the assignment process is said to be unbiased, and the resultant groups are equivalent.

Although the rationale underlying this principle is a bit complex, understanding how random assignment produces equivalent groups is a key point.

Farrington, Ohlin, and Wilson (1986:66) compare randomization in criminal justice research to laboratory controls in the natural sciences:

The control of extraneous variables by randomization is similar to the control of extraneous variables in the physical sciences by holding physical conditions (e.g., temperature, pressure) constant. Randomization insures that the average unit in [the] treatment group is approximately equivalent to the average unit in another [group] before the treatment is applied.

You've surely heard the expression "All other things being equal." Random assignment makes it possible to assume that all other things are equal.

Experiments and Causal Inference

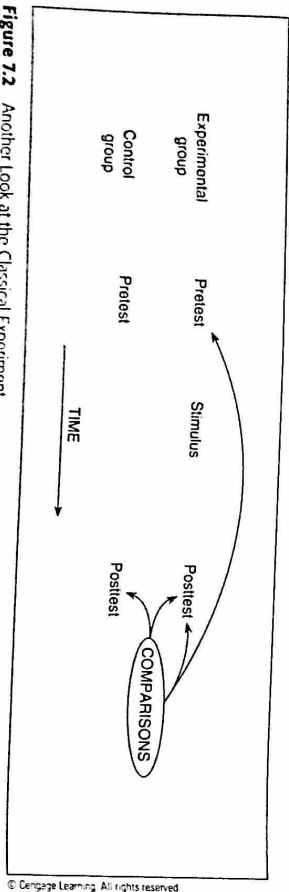
Experiments potentially control for many threats to the validity of causal inference, but researchers must remain aware of these threats.

The central features of the classical experiment are independent and dependent variables, pre-testing and posttesting, and experimental and control groups created through random assignment. Think of these features as *building blocks* of a

research design to demonstrate a cause-and-effect relationship. This point will become clearer by comparing the criteria for causality, discussed in Chapter 4, to the features of the classical experiment, as shown in Figure 7.2.

The experimental design ensures that the cause precedes the effect in time by taking posttest measurements of the dependent variable after introducing the experimental stimulus. The second criterion for causation—an empirical correlation between the cause-and-effect variables—is determined by comparing the pretest (in which the experimental stimulus is not present) to the posttest for the experimental group (after the experimental stimulus is administered). A change from pretest to posttest measures demonstrates correlation.

The final requirement is to show that the observed correlation between cause and effect is not due to the influence of some third variable. The classical experiment makes it possible to satisfy this criterion for cause in two ways. First, the posttest measures for the experimental group (stimulus present) are compared with those for the control group (stimulus not present). If the observed correlation between the stimulus and the dependent variable is due to some other factor, then the two posttest scores will be similar. Second, random assignment produces experimental and control groups that are equivalent and will not differ on some other variable that could account for the empirical correlation between cause and effect.



© Cengage Learning. All rights reserved.

Experiments and Threats to Validity

The classical experiment is designed to satisfy the three requirements for demonstrating cause-and-effect relationships. But what about threats to the validity of causal inference discussed in Chapter 4? In this section, we consider each of those threats in more detail and describe how the classical experiment reduces many of them. The book by Donald Campbell and Julian Stanley (1966) is the most frequently cited authority on threats to validity. For a fuller discussion, on which we draw heavily, see the book by William Shadish, Thomas Cook, and Donald Campbell (2002). We present these threats in a slightly different order, beginning with threats to internal validity.

Threats to Internal Validity

The problem of threats to internal validity refers to the possibility that conclusions drawn from experimental results may not accurately reflect what went on in the experiment itself. As we stated in Chapter 4, conclusions about cause and effect may be biased in some systematic way. Shadish, Cook, and Campbell (2002:54–60) pointed to several sources of the problem.

As you read about these different threats to internal validity, keep in mind that each is an example of a simple point: possible ways that researchers might be wrong in inferring causation. Although these threats are well known among researchers and often cited by students as things to memorize, they are simply examples of ways we might be wrong in causal inference.

History Historical events may occur during the course of the experiment that confound the experimental results. The hospitalization of a popular athlete for acute alcohol poisoning during an experiment on reducing alcohol use is an example.

Maturation People are continually growing and changing, whether in an experiment or not, and those changes affect the results of the experiment. In a long-term experiment, the fact that the subjects grow older (and wiser) may have an effect. In shorter experiments, subjects may become

tired, sleepy, bored, or hungry, or change in other ways that affect their behavior in the experiment. A long-term study of alcohol abuse might reveal a decline in binge drinking as the subjects mature.

History and maturation are similar in that they represent a correlation between cause and effect that is due to something other than the independent variable. They're different in that history represents something that's outside the experiment altogether, whereas maturation refers to change within the subjects themselves.

Testing Often, the process of testing and re-testing influences people's behavior and thereby confounds the experimental results. Suppose we administer a questionnaire to a group as a way of measuring their alcohol use. Then we administer an experimental stimulus and remeasure their alcohol use. By the time we conduct the posttest, the subjects may have become more sensitive to the issue of alcohol use and so may provide different answers. In fact, they may have figured out that we are trying to determine whether they drink too much. Because excessive drinking is frowned on by university authorities, our subjects will be on their best behavior and give answers that they think we want or that will make them look good.

Instrumentation Thus far, we haven't said much about the process of measurement in pre-testing and posttesting, and it's appropriate to keep in mind the problems of conceptualization and operationalization discussed in Chapter 5. If we use different measures of the dependent variable (say, different questionnaires about alcohol use), how can we be sure that they are comparable? Perhaps alcohol use seems to have decreased simply because the pretest measure was more sensitive than the posttest measure. Or if the measurements are being made by the experimenters, their procedures may change over the course of the experiment. That would be a problem of reliability.

Instrumentation is always a potential problem in criminal justice research that uses secondary sources of information such as police records about crime or court records about probation violations. There may be changes in how probation

violations are defined or changes in the recording practices of police departments.

The differences between testing and instrumentation threats to internal validity may seem unclear. In general, testing refers to changes in how subjects respond to measurement, whereas instrumentation is concerned with changes in the measurement process itself. If police officers respond differently to pretest and posttest questionnaires about prejudice, for example, that is a testing problem. However, if different questionnaires about prejudice are used in pretest and posttest measurements, instrumentation is a potential threat.

Statistical Regression Sometimes, it's appropriate to conduct experiments on subjects who start out with extreme scores on the dependent variable. A classic example is offered by sentencing policies that target chronic offenders. Charles Murray and L.A. Cox (1979) examined a program to incarcerate high-rate male juvenile offenders. They found that rearrest rates for subjects who had served sentences averaging 11 months were substantially lower than those for other offenders who were not incarcerated. These findings were questioned by Michael Maltz and associates (1980), who pointed out that the number of crimes committed by any given offender fluctuates over time. They argued that subjects in the experimental group—chronic offenders—were jailed following a period when their offense rates were abnormally high and that the decline in posttest arrests simply reflected a natural return to less extreme rates of offending. Even without any experimental stimulus, then, the group as a whole was likely to show some improvement over time.

Commonly referred to as “regression to the mean,” this threat to validity can emerge whenever researchers are interested in cases that have extreme scores on some variable. Consider hot spots, for example, which are areas where crime appears to be concentrated. Hot spots have become increasingly important in police operations. Michael Townsley and Ken Pease (2002) point out that it is important to consider the number of crimes in a hot spot and the time

period over which they are counted. It's common for temporary “spikes” in crime to appear in different areas of large cities, only to disappear in a couple of weeks without any action by police.

Selection Biases Random assignment eliminates the potential for systematic bias in selecting subjects. But subjects may be chosen in other ways that threaten validity. Volunteers are often solicited for experiments conducted on college campuses. Students who volunteer for an experiment may not be typical of students as a whole, however. Volunteers may be more interested in the subject of the experiment and more likely to respond to a stimulus. Or, if experimental subjects are paid some fee, students in greater financial need may participate, although they may not be representative of other students.

A common type of selection bias in applied criminal justice studies results from the natural caution of public officials. Let's say you are a bail commissioner in a large city, and the mayor wants to try a new program to increase the number of arrested persons who are released on bail. The mayor asks you to determine what kinds of defendants should be eligible for release and informs you that staff from the city's criminal justice services agency will be evaluating the program. In establishing eligibility criteria, you will probably try to select defendants who will not be arrested again while on bail and defendants who will most likely show up for scheduled court appearances. In other words, you will try to select participants who are least likely to fail. This common and understandable caution is sometimes referred to as “creaming”—skimming the best risks off the top. Creaming is a threat to validity because the low-risk persons selected for release, although most likely to succeed, do not represent the jail population as a whole.

Experimental Mortality Experimental subjects often drop out of an experiment before it is completed, and that can affect statistical comparisons and conclusions. This is termed “experimental mortality,” sometimes referred to as attrition. In the classical experiment involving an experimental and a control group, each with a pretest

and a posttest, suppose that the heavy drinkers in the experimental group are so turned off by the video on the health effects of binge drinking that they tell the experimenter to forget it and leave. Those subjects who stick around for the posttest are less heavy drinkers to start with, and the group results will thus reflect a substantial “decrease” in alcohol use.

In this example, mortality is related to the experimental stimulus itself: subjects who score highest on the pretest are the most likely to drop out after viewing the video. Mortality may also be a problem in experiments that take place over a long period (people may move away) or in experiments that require a substantial commitment of effort or time by subjects, they may become bored with the study or simply decide it's not worth the effort.

Causal Time Order In criminal justice research, there may be ambiguity about the time order of the experimental stimulus and the dependent variable. Whenever this occurs, the research conclusion that the stimulus caused the dependent variable can be challenged with the explanation that the “dependent” variable actually caused changes in the stimulus. Many early studies of the relationship between different types of punishments and rates of offending exhibited this threat to validity by relying on single interviews with subjects who were asked how they viewed alternative punishments and whether they had committed any crimes.

Diffusion or Imitation of Treatments In the event that experimental- and control-group subjects are in communication with each other, it's possible that experimental subjects will pass on some elements of the experimental stimulus to the control group. David Weisburd, Nancy A. Morris, and Justin Ready (2008:193) describe the potential for spillover effects in their evaluation of targeted community policing in Redlands, California:

While experimental treatments were restricted to the experimental block groups and juveniles living within them, it is reasonable to expect some degree of diffusion of the treatment

effects. One source of that diffusion may have been the Redlands [sic] schools.... While treatment in the schools was focused only on students in the experimental areas, it is clear that experimental and control students in the same schools would likely interact and pass on information.

Any program that provides a treatment only to some people in a school or other institution and not to others in the same institution is potentially vulnerable to such spillover effects.

Compensatory Treatment In many criminal justice experiments—such as a special job-training program for incarcerated felons—subjects in the control group may be deprived of something of value to them. In such cases, there may be pressures to offer some form of compensation. Recall the discussion in Chapter 4 of how police in the Kansas City Preventive Patrol Experiment controlled the perimeter of reactive beats—those with no preventive patrol. As we noted, they more often used lights and sirens when responding to calls for service in the reactive beats. Some police officers compensated for the absence of preventive patrol in a way that reduced the differences among pro-active, reactive, and control beats.

It's helpful to think of treatment diffusion as the accidental spillover of an experimental stimulus, in contrast to the more intentional compensatory treatment by public officials that occurred in the Kansas City case. In applied criminal justice research, such treatment is probably more common, and researchers often take steps to prevent it. For example, in an evaluation of intensive supervision probation (ISP) programs in 11 sites, researchers from the RAND Corporation recognized that probation officers might provide enhanced supervision to clients in the control group, under the assumption that the experimental program (ISP) was better than traditional probation (Peterson, 1989). This potential for compensation was reduced in two ways. First, researchers explained to program staff why the experimental and control groups should receive different types of services. Second, a type of double-blind procedure was used whereby researchers tried to disguise the

records of subjects in the control group so that probation staff could not distinguish subjects who were participating in the experiment from their regular probation caseload.

Compensatory Rivalry In real-life experiments, subjects deprived of the experimental stimulus may try to compensate by working harder. Suppose an experimental career development program for corrections officers is the experimental stimulus; the control group may work harder than before in an attempt to keep pace with the "special" experimental subjects.

Demoralization Feelings of deprivation among the control group may also result in subjects giving up. A career development program for corrections officers may prompt demoralization among control-group subjects who believe their opportunities for advancement will suffer. As a result, it may not be clear whether posttest differences between experimental and control groups in, say, job performance actually reflected program impacts or were due to demoralization among the control group.

Notice that the possibilities of compensatory rivalry and demoralization are based on subjects' reactions to the experiment. Diffusion and compensatory treatment are accidental or intentional extensions of the experimental stimulus to the control group. In studies in which agency staff, not researchers, administer an experimental treatment to subjects, there is a greater potential for intentional compensation. In studies in which subjects in a control group are aware that other subjects are receiving a desirable treatment, compensatory rivalry or demoralization is possible.

Ruling Out Threats to Internal Validity

There, then, are the threats to internal validity cited by Shadish, Cook, and Campbell. The classical experiment, coupled with proper subject selection and assignment, can potentially handle each of the 12 threats to internal validity.

How do researchers determine whether a particular design rules out threats to internal validity? Shadish, Cook, and Campbell provide excellent

advice. Ruling out threats to validity requires a "theory of plausibility" so that we know which of the many possible threats are plausible in [a particular context] (2002:41). In other words, some threats make more sense in some settings than in others. Sorting this out requires having theory-based expectations to guide us in figuring out what validity threats might be at work in a particular situation. Let's look again at the classical experiment, presented graphically in Figure 7.2.

Pursuing the example of the educational video as an attempt to reduce alcohol abuse, if we use the experimental design shown in Figure 7.2, we should expect two findings. For the experimental group, the frequency of drinking measured in their posttest should be less than in their pretest. In addition, when the two posttests are compared, the experimental group should have less drinking than the control group.

This design guards against the problem of history because anything occurring outside the experiment that might affect the experimental group should also affect the control group. There should still be a difference in the two posttest results. The same comparison guards against problems of maturation as long as the subjects have been randomly assigned to the two groups. Testing and instrumentation should not be problems because both the experimental and the control groups are subject to the same tests and experimenter effects. If the subjects have been assigned to the two groups randomly, statistical regression should affect both equally, even if people with extreme scores on drinking (or whatever the dependent variable is) are being studied. Selection bias is ruled out by the random assignment of subjects.

Experimental mortality can be more complicated to handle because dropout rates may be different between the experimental and control groups. The experimental treatment itself may increase mortality in the group exposed to the video. As a result, the group of experimental subjects that received the posttest will differ from the group that received the pretest. In our example of the alcohol video, it would probably not be possible to handle this problem by administering a placebo, for instance. In general, however, the

potential for mortality can be reduced by shortening the time between pretest and posttest, by emphasizing to subjects the importance of completing the posttest, or perhaps by offering cash payments for participating in all phases of the experiment.

The remaining problems of internal invalidity can be avoided through the careful administration of a controlled experimental design. The experimental design we've been discussing facilitates the clear specification of independent and dependent variables. Experimental and control subjects can be kept separate to reduce the possibility of diffusion or imitation of treatments. Administrative controls can be applied to avoid compensations given to the control group, and compensatory rivalry can be watched for and taken into account in evaluating the results of the experiment, as can the problem of demoralization.

We emphasize careful administration here. Random assignment, pretest and posttest measures, and use of control and experimental groups do not automatically rule out threats to validity. This caution is especially true in field studies and evaluation research, in which subjects participate in natural settings and uncontrolled variation in the experimental stimulus may be present. Control over experimental conditions is the hallmark of this approach, but conditions in field settings are usually more difficult to control.

Compare, for example, our hypothetical study of alcohol use among college students with the field experiment on ISP described by Joan Petersilia (1989). In Petersilia's study, more intensive probation was the independent variable, and recidivism was the dependent variable. Subjects were randomly assigned to the experimental group (ISP) or the control group (regular probation).

The alcohol use study could conceivably be completed in about 1 week, using subjects from a class, dormitory, or house on a single campus. ISP programs were evaluated in 11 sites over 4 years, using probation clients as subjects. The video on the health effects of alcohol use is a well-defined treatment that is readily standardized and can easily be controlled by researchers. The experimental treatment in the ISP programs was a reduced

caseload for probation workers, together with an increased number of regular contacts with each probation client; program staff, not researchers, administered the experimental treatment. There was a great potential for uncontrolled variation in the delivery of ISP treatments; ISP is not a simple dichotomous treatment as is the video/no-video treatment in the alcohol use experiment.

Finally, the alcohol use questionnaire can easily be administered by researchers, providing reliable measures of alcohol use. Data on recidivism for the ISP study were collected by probation staff in each of the 11 sites. Although recidivism can be readily defined as the number of new arrests after beginning probation, there may have been wide variation in the ability of staff in the 11 sites to reliably detect new arrests.

These remarks are not intended as criticism of the ISP study. In her description of the evaluation, Petersilia (1989) documents the extensive steps taken by RAND Corporation researchers to control possible validity threats. The important point is that field experiments and evaluations can present many obstacles that are not eliminated simply by adopting a randomized experimental design. Careful administration and control throughout the experiment are necessary to reduce potential threats to internal validity.

Generalizability and Threats to Validity

Potential threats to internal validity are only some of the complications faced by experimenters. They also have the problem of generalizing from experimental findings to the real world. Even if the results of an experiment are an accurate gauge of what happened during that experiment, do they really tell us anything about life in the wilds of society? With our examination of cause and effect in Chapter 4 in mind, we consider two dimensions of **generalizability**: construct validity and external validity.

Threats to Construct Validity

In the language of experimentation, construct validity is the correspondence between the empirical test of a hypothesis and the underlying causal

process that the experiment is intended to represent. Construct validity is thus concerned with generalizing from our observations in an experiment to actual causal processes in the real world. In our hypothetical example, the educational video is how we operationalize the construct of understanding the health effects of alcohol abuse. Our questionnaire represents the dependent construct of actual alcohol use.

Are these reasonable ways to represent the underlying causal process in which understanding the effects of alcohol use causes people to reduce excessive or abusive drinking? It is a reasonable representation, but also one that is certainly incomplete. People develop an understanding of the health effects of alcohol use in many ways. Watching an educational video is one way; having personal experience, talking to friends and parents, taking other courses, and reading books and articles are others. Our video may do a good job of representing the health effects of alcohol use, but it is an incomplete representation of that construct. Alternatively, the video may be poorly produced, too technical, or incomplete. Then the experimental stimulus may not adequately represent the construct we are interested in—educating students about the health effects of alcohol use. There may also be problems with our measure of the dependent variable: questionnaire items on self-reported alcohol use.

Here's another example from the popular author Malcolm Gladwell (2003). Part of his book, *Line: The Power of Thinking Without Thinking*, describes how market researchers conduct experiments to test new consumer products. In blind tests of cola drinks, participants were given small samples of different colas without knowing the cola's brand. Market researchers found that people preferred a sweeter type of cola in these experiments. However, in follow-up tests where participants drank 12-ounce servings, more preferred the less sweet formula. A small taste—2 ounces or so—did not represent the construct of cola drinking for most people, who typically drink about 12 ounces. In smaller doses people liked sweeter cola, but that was an artifact of the experiment (2003:165–166).

By this time, you should recognize a similarity between construct validity and some of the measurement issues discussed in Chapter 5. Almost any empirical example or measure of a construct is incomplete. Part of construct validity involves how completely an empirical measure can represent a construct or how well we can generalize from a measure to a construct.

A related issue in construct validity is whether a given level of treatment is sufficient. Perhaps showing a single video to a group of subjects would have little effect on alcohol use, but administering a series of videos over several weeks would have a greater impact. We could test this experimentally by having more than one experimental group and varying the number of videos seen by different groups.

Threats to construct validity are problematic in criminal justice experiments, often because researchers do not clearly specify precisely what constructs are to be represented by particular measures or experimental treatments. Farrington, Ohlin, and Wilson (1986:92) make a related point: "Most treatments in existing experiments are not based on a well-developed theory but on a vague idea about what might influence offending. The treatments given are often heterogeneous, making it difficult to know which element was responsible for any observed effect." These authors also note the importance of thinking about levels of constructs.

The RAND Corporation evaluation of intensive supervision provides a good example. The ISP program explicitly defined enhanced probation as punitive. This leads to the question of how much probation should be enhanced. If a typical probation officer carries an average caseload of 100 clients and sees each an average of twice per month, what workload and contact levels are sufficiently intensive to reduce recidivism? Cutting the workload in half and doubling the number of contacts would be more intensive, but would it be intensive enough to produce a decline in recidivism? Again, an experiment could test this question by including more than one experimental group and giving each a different level of probation supervision.

In summary, three elements in enhancing construct validity are (1) linking constructs and measures to theory, (2) clearly indicating what constructs are represented by specific measures, and (3) thinking carefully about what levels of treatment may be necessary to produce some level of change in the dependent measure.

Threats to External Validity

Will an experimental study, conducted with the kind of control we have emphasized here, produce results that would also be found in more natural settings? Can an intensive probation program shown to be successful in Minneapolis achieve similar results in Miami? External validity represents a slightly different form of generalizability, one in which the question is whether results from experiments in one setting (time and place) will be obtained in other settings, or whether a treatment found to be effective for one population will have similar effects on a different group.

Threats to external validity are greater for experiments conducted under carefully controlled conditions. If the alcohol education experiment reveals that drinking decreased among students in the experimental group, then we can be confident that the video actually reduced alcohol use among our experimental subjects. But will the video have the same effect on high school students or adults if it is broadcast on television? We cannot be certain because the carefully controlled conditions of the experiment might have had something to do with the video's effectiveness.

In contrast, criminal justice field experiments are conducted in more natural settings. Real probation officers in 11 different local jurisdictions delivered intensive supervision to real probationers in the RAND ISP experiment. Because of the real-world conditions and multiple sites, there were fewer potential threats to external validity. This is not to say that external validity is never a problem in field experiments. The 11 probation agencies that participated in this evaluation may not be typical of probation agencies in other areas. The simple fact that they were willing to participate in the study suggests that staff could be

more dedicated or more amenable to trying new approaches to probation. One of the advantages of field experiments in criminal justice is that, because they take place under real-world conditions, results are more likely to be valid in other real-world settings as well.

You may have detected a fundamental conflict between internal and external validity. Threats to internal validity are reduced by conducting experiments under carefully controlled conditions. But such conditions do not reflect real-world settings, and this restricts our ability to generalize results. Field experiments generally have greater external validity, but their internal validity may suffer because such studies are more difficult to monitor than those taking place in more controlled settings. This is what John Eck (2002:104) refers to as a "diabolical dilemma."

Shadish, Cook, and Campbell (2002:101) offer some useful advice for resolving the potential for conflict between internal and external validity. Explanatory studies that test cause-and-effect theories should place greater emphasis on internal validity; applied studies should be more concerned with external validity. This is not a hard-and-fast rule because internal validity must be established before external validity becomes an issue. That is, applied researchers must have confidence in the internal validity of their cause-and-effect relationships before they ask whether similar relationships would be found in other settings.

Threats to Statistical Conclusion Validity

The basic principle of statistical conclusion validity is simple. Virtually all experimental research in criminal justice is based on samples of subjects that represent a target population. Larger samples of subjects, up to a point, are more representative of the target population than are smaller samples. Statistical conclusion validity becomes an issue when findings are based on small samples of cases. Because experiments can be costly and time consuming, they are frequently conducted with relatively small numbers of subjects. In such cases, only large differences between experimental and control groups on posttest

measures can be detected with any degree of confidence.

In practice, this means that finding cause-and-effect relationships through experiments depends on two related factors: (1) the number of subjects and (2) the magnitude of posttest differences between the experimental and control groups. Experiments with large numbers of cases may be able to reliably detect small differences, but experiments with smaller numbers can detect only large differences.

Threats to statistical conclusion validity can be magnified by other difficulties in field experiments. If treatment spillover or compensation is a problem, then smaller differences in the experimental stimulus will be delivered to each group. More generally, David Weisburd, Cynthia Lum, and Sue-Ming Yang (2003) concluded, after reviewing a large number of criminal justice experiments, that failure to maintain control over experimental conditions reduces statistical conclusion validity even for studies with large numbers of subjects. Furthermore, these authors found that as sample size increases, so do implementation difficulties, which undermines experimental results in a variety of ways.

Variations in the Classical Experimental Design

The basic experimental design is adapted to meet different research applications.

We now turn to a more systematic consideration of variations on the classical experiment that can be produced by manipulating the building blocks of experiments.

Slightly restating our earlier remarks, four basic building blocks are present in experimental designs: (1) the number of experimental and control groups, (2) the number and variation of experimental stimuli, (3) the number of pretest and posttest measurements, and (4) the procedures used to select subjects and assign them to groups. By way of illustrating these building blocks and the ways they are used to produce different

designs, we adopt the system of notation used by Campbell and Stanley (1966). Figure 7.3 presents this notation and shows how it is used to represent the classical experiment and examples of variations on this design.

In Figure 7.3, the letter *O* represents observations or measurements, and *X* represents an experimental stimulus or treatment. Different time points are displayed as *t* with a subscript to represent time order. Thus, for the classical experiment shown in Figure 7.3, *O* at *t*₁ is the pretest, *O* at *t*₂ is the posttest, and the experimental stimulus, *X*, is administered to the experimental group at *t*₂ between the pretest and posttest. Measures are taken for the control group at times *t*₁ and *t*₂, but the experimental stimulus is not administered to the control group.

Now consider the design labeled "Posttest Only." As implied by its name, no pretest measures are made on either the experimental or the control group. Thinking for a moment about the threats to internal validity, we can imagine situations in which a posttest-only design is appropriate. Testing and retesting might especially influence

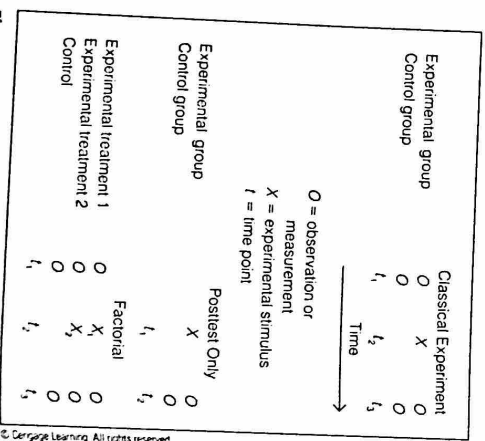


Figure 7.3 Variations in the Experimental Design

subjects' behavior if measurements are made by administering a questionnaire, with subjects' responses to the posttest potentially affected by their experience in the pretest. A posttest-only design can reduce the possibility of testing as a threat to validity by eliminating the pretest.

Without a pretest, it is obviously not possible to detect change in measures of the dependent variable, but we can still test the effects of the experimental stimulus by comparing posttest measures for the experimental group with posttest measures for the control group. For example, if we are concerned about the possibility of sensitizing subjects in a study of an alcohol education video, we might eliminate the pretest and examine the posttest differences between the experimental and control groups. Random assignment is the key to the posttest-only design. If subjects are randomly assigned to experimental and control groups, we expect them to be equivalent. Any posttest differences between the two groups on the dependent variable can then be attributed to the influence of the video.

In general, posttest-only designs are appropriate when researchers suspect that the process of measurement may bias subjects' responses to a questionnaire or other instrument. This is more likely when only a short time elapses between pretest and posttest measurements. The number of observations made on subjects is a design building block that can be varied as needed. We emphasize here that random assignment is essential in a posttest-only design.

Figure 7.3 also shows a factorial design, which has two experimental groups that receive different treatments, or different levels of a single treatment, and one control group. This design is useful for comparing the effects of different interventions or different amounts of a single treatment. In evaluating an ISP program, we might wish to compare how different levels of contact between probation officers and probation clients affect recidivism. In this case, subjects in one experimental group might receive weekly contact (*X*₁), subjects in the other experimental group might be seen by probation officers twice each week (*X*₂), and control-group subjects might have normal

contact (say, monthly) with probation officers. Because more contact is more expensive than less contact is, we would be interested in seeing how much difference in recidivism was produced by monthly, weekly, and twice-weekly contacts.

Thus, an experimental design may have more than one group receiving different versions or levels of experimental treatment. We can also vary the number of measurements made on dependent variables. No hard-and-fast rules exist for using these building blocks to design a given experiment. A useful rule of thumb, however, is to keep the design as simple as possible to control for potential threats to validity. The specific design for any particular study depends on the research purpose, available resources, and unavoidable constraints in designing and actually carrying out the experiment.

One very common constraint is how subjects or units of analysis are selected and assigned to experimental or control groups. This building block brings us to the subject of quasi-experimental designs.

Quasi-Experimental Designs

When random assignment is not possible, researchers can use different types of quasi-experimental designs.

By now, the value of random assignment in controlling threats to validity should be apparent. However, it is often impossible to randomly select subjects for experimental and control groups and satisfy other requirements. Most often, there may be practical or administrative obstacles. There may also be legal or ethical reasons why random assignment cannot be used in criminal justice experiments.

When random assignment is not possible, the next-best choice is often a **quasi-experiment**. The prefix *quasi*, meaning "to a certain degree,"

Quasi-experiment Research design that includes most, but not all, elements of an experiment.

is significant—a quasi-experiment is, to a certain degree, an experiment. In most cases, quasi-experiments do not randomly assign subjects and, therefore, may suffer from the internal validity threats that are so well controlled in true experiments. Without random assignment, the other building blocks of experimental design must be used creatively to reduce validity threats. Following Shadish, Cook, and Campbell (2002), we will group quasi-experimental designs into two categories: (1) nonequivalent-groups designs and (2) time-series designs. Each can be represented with the same O , X , and t notation used to depict experimental designs.

Nonequivalent-Groups Designs

The name for this family of designs is also meaningful. The main strength of random assignment is that it allows us to assume equivalence in experimental and control groups. When it is not possible to create groups through random assignment, we must use some other procedure, one that is not random. If we construct groups through some nonrandom procedure, however, we cannot assume that the groups are equivalent—hence, the label *nonequivalent-groups design*.

Whenever experimental and control groups are not equivalent, we should select subjects in some way that makes the two groups as comparable as possible. Often, the best way to achieve comparability is through a matching process in which subjects in the experimental group are matched with subjects in a comparison group. The term *comparison group* is commonly used, rather than *control group*, to highlight the nonequivalence of groups in quasi-experimental designs. A comparison group does, however, serve the same function as a control group.

Some examples of research that uses nonequivalent-groups designs illustrate various approaches to matching and the creative use of experimental design building blocks. Examples include studies of child abuse (Widom, 1989a), obscene phone calls (Clarke, 1997a), and video cameras for crime prevention (Gill and Spriggs, 2005). Figure 7.4 shows a diagram of each design using the X , O , and t notation. The solid line that separates treatment and comparison groups in the figure signifies that

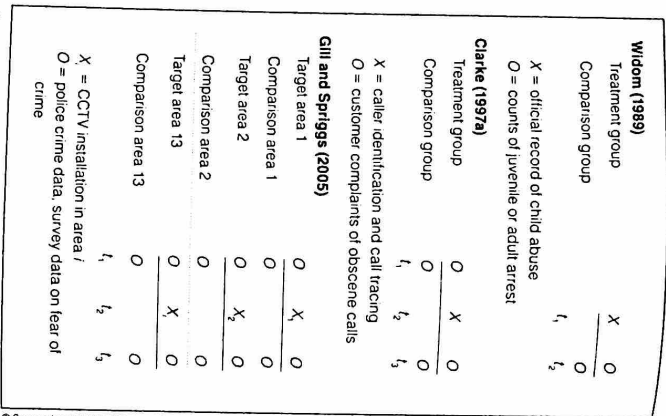


Figure 7.4 Quasi-Experimental Design Examples

subjects have been placed in groups through some nonrandom procedure.

Child Abuse and Later Arrest Cathy Spatz Widom (1989a; also Widom and Maxfield, 2001) studied the long-term effects of child abuse—whether abused children are more likely to be charged with delinquent or adult criminal offenses than are children who were not abused. Child abuse was the experimental stimulus, and the number of subsequent arrests was the dependent variable.

Of course, it is not possible to assign children randomly to groups in which some are abused and others are not. Widom's design called for selecting a sample of children who, according to court records, had been abused. She then matched each abused subject with a comparison subject—of the

same gender, race, age, and approximate socioeconomic status (SES)—who had not been abused. The assumption with these matching criteria was that age at the time of abuse, gender, race, and SES differences might confound any observed relationship between abuse and subsequent arrests. Widom found that children who had been abused or neglected were more likely to have juvenile records or adult arrests. But the differences between the treatment and comparison groups were smaller than differences found in most previous research.

You may be wondering how a researcher selects important variables to use in matching experimental and comparison subjects. We cannot provide a definitive answer to that question, any more than we can specify what particular variables should be used in a given experiment. The answer ultimately depends on the nature and purpose of the experiment. As a general rule, however, the two groups should be comparable in terms of variables that are likely to be related to the dependent variable under study. Widom matched on gender, race, and SES because these variables are correlated with juvenile and adult arrest rates. Age at the time of reported abuse was also an important variable because children abused at a younger age had a longer "at-risk" period for delinquent arrests.

Widom produced experimental and comparison groups by matching individual subjects. It is also possible to construct experimental and comparison groups through aggregate matching, in which the average characteristics of each group are comparable. This is illustrated in our next example.

Detering Obscene Phone Calls In 1988, the New Jersey Bell telephone company introduced caller identification (ID) and instant call tracing in a small number of telephone exchange areas. Now ubiquitous in mobile phones, caller ID was a new technology in 1988. Instant call tracing allows the recipient of an obscene or threatening call to automatically initiate a procedure to trace the source of the call.

Ronald Clarke (1997a) studied the effects of these new technologies in deterring obscene phone calls. Clarke expected that obscene calls would decrease in areas where the new services were

available. To test this, he compared records of formal customer complaints about annoying calls in the New Jersey areas that had the new services to formal complaints in other New Jersey areas where caller ID and call tracing were not available. One year later, the number of formal complaints had dropped sharply in areas serviced by the new technology; no decline was found in other New Jersey Bell areas.

In this study, telephone service areas with new services were the treatment group, and areas without the services were the comparison group. Clarke's matching criterion was a simple one: telephone service by New Jersey Bell, assuming the volume of obscene phone calls was relatively constant within a single phone service area. Of course, matching on telephone service area cannot eliminate the possibility that the volume of obscene phone calls varies from one part of New Jersey to another, but Clarke's choice of a comparison group was straightforward and certainly more plausible than comparing New Jersey to, say, New Mexico.

Clarke's study is a good example of a natural field experiment. The experimental stimulus—caller ID and call tracing—was not specifically introduced by Clarke, but he was able to obtain measures for the dependent variable before and after the experimental stimulus was introduced. This design made it possible for Clarke to call with reasonable confidence that caller ID and call tracing reduced the number of formal complaints about obscene phone calls.

Another useful lesson from this study is related to the dramatic drop in obscene phone calls. Donald Campbell, whose name has long been associated with experimentation in social science, argues convincingly that researchers can have more confidence in the validity of large changes in dependent variables compared with small changes (Campbell, 1979). Ronald Clarke (2008) refers to this as a "cliff-edge" effect, meaning that some crime prevention measures produce reductions in crime that figuratively fall off the cliff.

Cameras and Crime Prevention U.S. residents have become accustomed to seeing closed-circuit television (CCTV) cameras in stores and at ATMs, but this technology is less used in public spaces

such as streets and parking lots. With an estimated 4 million cameras deployed, CCTV is widely used as a crime prevention and surveillance tool in the United Kingdom (McCaill and Norris, 2003). CCTV enabled the London Metropolitan Police to quickly identify suspects in the Underground bombing attacks that took place in 2005. Cameras are increasingly used to monitor traffic and even record license plates of cars running traffic lights. But does CCTV have any effect in reducing crime?

Martin Gill and Angela Spriggs (2005) conducted an evaluation of 13 CCTV projects installed in a variety of residential and commercial settings in England. These were a mix of small- and large-scale CCTV projects involving multiple cameras. One area on the outskirts of London included over 500 cameras installed to reduce thefts of and from vehicles in parking facilities. Five projects in London and other urban areas placed 10–15 cameras in low-income housing areas, seeking to reduce burglary and robbery. Researchers examined two types of dependent variables before and after cameras were installed: crimes reported to police and fear of crime. Fear was measured through surveys of people living in residential areas and samples of people living in residential areas and samples of people on local streets for commercial areas and parking facilities.

Measuring police data and fear of crime before and after cameras were installed made it possible for Gill and associates to satisfy two criteria for cause–time order and covariation between the independent variable (CCTV) and dependent variables. Random assignment of areas to treatment (CCTV) or control (no CCTV) groups was not possible. This was because the intervention was planned for only a small number of locations of each type (residential, commercial, parking) and because CCTV was carefully tailored to each site. Instead, the researchers created two types of comparison areas. First, comparison areas “were selected by similarity on socio-demographic and geographical characteristics and crime problems” (Gill and Spriggs, 2005:123–124). The second type of comparison was “buffer zones,” defined as an area in a 1-mile radius from the edge of the target area where CCTV cameras were installed; buffer zones were defined only for CCTV areas.

The rationale for comparison areas is clear. If CCTV is effective in reducing crime, we should

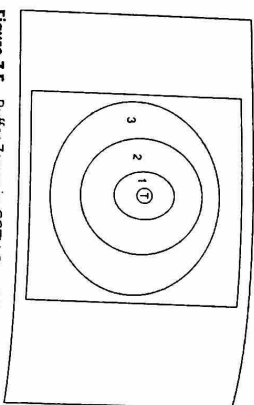


Figure 7.5 Buffer Zones in CCTV Quasi-Experiment
Source: Adapted from Gill and Spriggs (2005:40).

expect declines in target areas, but not in comparison areas. Alternatively, if posttreatment measures of crime went down in both treatment and comparison areas, we might expect greater declines in the CCTV sites. But what about buffer areas? After defining buffer areas, researchers then subdivided them into concentric rings around a target area (1), as shown in Figure 7.5. The stated purpose was to assess any movement of crime around the target area. If CCTV was effective in reducing crime, any reduction should be greatest in the target area; the size of the reduction should decline moving outward from the target area.¹

Short-term results found some reduction of some types of crime in some CCTV areas. In other treatment areas, some crimes increased more than in comparison areas. In particular, Gill and Spriggs (2005) found that public order offenses such as drunkenness tended to increase more in CCTV target areas. Overall, significant drops in crime were found in just 2 of 13 target areas. Fear and related attitudes declined in all target and comparison areas, but the authors believed this was largely due to declining crime in all areas.

This example illustrates why nonequivalent comparison groups are important. Because crime declined in most areas and fear declined in all, a simple comparison of pre- and postintervention measures would have been misleading. That strategy would have suggested that CCTV was responsible for reduced crime and fear. Only by

adding the comparison and buffer areas to their research were Gill and Spriggs (2005) able to learn that CCTV was probably not the cause of declines because similar patterns were found in many areas where CCTV systems were not installed.

Together, these three studies illustrate different approaches to research design when it is not possible to randomly assign subjects to treatment and control groups. Lacking random assignment, researchers must use creative procedures for selecting subjects, constructing treatment and comparison groups, measuring dependent variables, and exercising other controls to reduce possible threats to validity.

Cohort Designs

Chapter 4 mentioned cohort studies as examples of longitudinal designs. We can also view cohort studies as nonequivalent-control-group designs. Recall from Chapter 4 that a cohort may be defined as a group of subjects who enter or leave an institution at the same time. For example, a class of police officers who graduate from a training academy at the same time could be considered a cohort. Or we might view all persons who were sentenced to probation in May as a cohort.

Now think of a cohort that is exposed to some experimental stimulus. The May probation cohort might be required to complete 100 hours of community service in addition to meeting other conditions of probation. If we are interested in whether probationers who receive community service sentences are charged with fewer probation violations, we can compare the performance of the May cohort with that of the April cohort or the June cohort or some other cohort not sentenced to community service. Cohorts that do not receive community service sentences serve as comparison groups. The groups are not equivalent because they were not created by random assignment. But if we assume that a comparison cohort does not systematically differ from a treatment cohort on important variables, we can use this design to determine whether community service sentences reduce probation violations.

That last assumption is very important, but it may not be viable. Perhaps a criminal court docket

is organized to schedule certain types of cases at the same time, so a May cohort would be systematically different from a June cohort. But if the assumption of comparability can be met, cohorts may be used to construct nonequivalent comparison and experimental groups by taking advantage of the natural flow of cases through some institutional process.

A study by Edmund Ristler, Tim Swearman, and Larry Nackerud (1998) on the effectiveness of a judicial waiver law in deterring juvenile crime is a good example of a cohort design. Judicial waiver laws signaled a shift in juvenile justice philosophy, largely because of a perceived increase in violent crime by juveniles and a growing sentiment that juvenile courts were not tough enough on young violent offenders. A judicial waiver law in Georgia required that anyone between the ages of 13 and 17 be tried as an adult if charged with one of several violent crimes: murder, voluntary manslaughter, rape, aggravated sexual battery, aggravated child molestation, aggravated sodomy, or robbery with a firearm. Ristler and associates were interested in whether the law achieved its goal of deterrence.

The Georgia law—the experimental stimulus—became effective in 1994. Two samples of juveniles were studied: (1) a preintervention cohort, juveniles processed during the 2-year period prior to the law's enactment (1992–1993); and (2) a postintervention cohort, juveniles processed during the 2-year period after the law took effect (1994–1995). Mean arrest rates for juveniles for the seven offenses covered under the judicial waiver law were compared across these two samples. Arrest rates for 1992 and 1993 served as observations for the pretest cohort; those for 1994 and 1995 served as posttest observations.

The two cohorts could not be considered equivalent because they were not produced by random assignment. However, they were assumed to be sufficiently similar for the purpose of the authors' study.

Ristler and associates found that the mean arrest rate for crimes covered under the judicial waiver law actually increased slightly for the postintervention cohort. The authors therefore concluded that the law did not act as a deterrent to reduce violent offending by juveniles.

¹Compare this with the Chicago research by Shaw and McKay described in Chapter 2.

Time-Series Designs

Time-series designs are common examples of longitudinal studies in criminal justice research. As the name implies, a time-series design involves examining a series of observations on some variable over time. A simple example is examining trends in arrests for drunk driving over time to see whether the number of arrests is increasing, decreasing, or staying constant. A police executive might be interested in keeping track of arrests for drunk driving, or for other offenses, as a way of monitoring the performance of patrol officers. Or state corrections officials might want to study trends in prison admissions as a way of predicting the future need for correctional facilities.

An interrupted time series is a special type of time-series design that can be used in cause-and-effect studies. A series of observations is compared before and after some intervention is introduced. For example, a researcher might want to know whether roadside sobriety checkpoints cause a decrease in fatal automobile accidents. Trends in accidents could be compared before and after the roadside checkpoints are established.

Interrupted time-series designs can be very useful in criminal justice research, especially in applied studies. They do have some limitations, however, just like other ways of structuring research. Shadish, Cook, and Campbell (2002) describe the strengths and limitations of different approaches to time-series designs. We will introduce these approaches with a hypothetical example and then describe some specific criminal justice applications. Continuing with the example of sobriety checkpoints, figure 7.6 presents four possible patterns of alcohol-related automobile accidents. The vertical line in each pattern shows the time when the roadside checkpoint program is introduced. Which of these patterns indicates that the new program caused a reduction in car accidents?

If the time-series results looked like pattern 1 in figure 7.6, we might think initially that the checkpoints caused a reduction in alcohol-related accidents, but there seems to be a general downward trend in accidents that continues after the

intervention. It's safer to conclude that the decline would have continued even without the roadside checkpoints.

Pattern 2 shows that an increasing trend in auto accidents has been reversed after the intervention, but this appears to be due to a regular pattern in which accidents have been bouncing up and down. The intervention was introduced at the peak of an upward trend, and the later decline may be an artifact of the underlying pattern rather than of the new program.

Patterns 1 and 2 exhibit some outside trend, rather than an intervention, that may account for a pattern observed over time. We may recognize this as an example of history as a validity threat to the inference that the new checkpoint program caused a change in auto accidents. The general decline in pattern 1 may be due to reduced drunk driving that has nothing to do with sobriety checkpoints. Pattern 2 illustrates what is referred to as seasonality in a time series—a regular pattern of change over time. Shadish, Cook, and Campbell (2002) describe seasonality as a special case of history. In our example, the data might reflect seasonal variation in alcohol-related accidents that occur around holidays or maybe on football weekends near a college campus.

Patterns 3 and 4 lend more support to the inference that sobriety checkpoints caused a decline in alcohol-related accidents, but the two patterns are different in a subtle way. In pattern 3, accidents decline more sharply from a general downward trend immediately after the checkpoint program was introduced, whereas pattern 4 displays a sharper decline some time after the new program was established. Which pattern provides stronger support for the inference?

In framing your answer, recall what we have said about construct validity. Think about the underlying causal process these two patterns represent, or consider possible mechanisms that might be at work. Pattern 3 suggests that the program was immediately effective and supports what we might call an incapacitation mechanism: roadside checkpoints enabled police to identify and arrest drunk drivers, thereby getting them off the road and reducing accidents. Pattern 4 suggests a

deterrent mechanism: as drivers learned about the checkpoints, they less often drove after drinking, and accidents eventually declined. Either explanation is possible given the evidence presented. This illustrates an important limitation of interrupted time-series designs: they operationalize complex causal constructs in simple ways. Our interpretation depends in large part on how we understand this causal process.

CHAPTER 7 Experimental and Quasi-Experimental Designs 189

The classic study by Richard McCleary, Barbara Nienstedt, and James Erven (1982) illustrates the need to think carefully about how well time-series results reflect underlying causal patterns. Recall that McCleary and colleagues reported a sharp decline in burglaries immediately after a special burglary investigation unit was established in a large city. This finding was at odds with their understanding of how police investigations

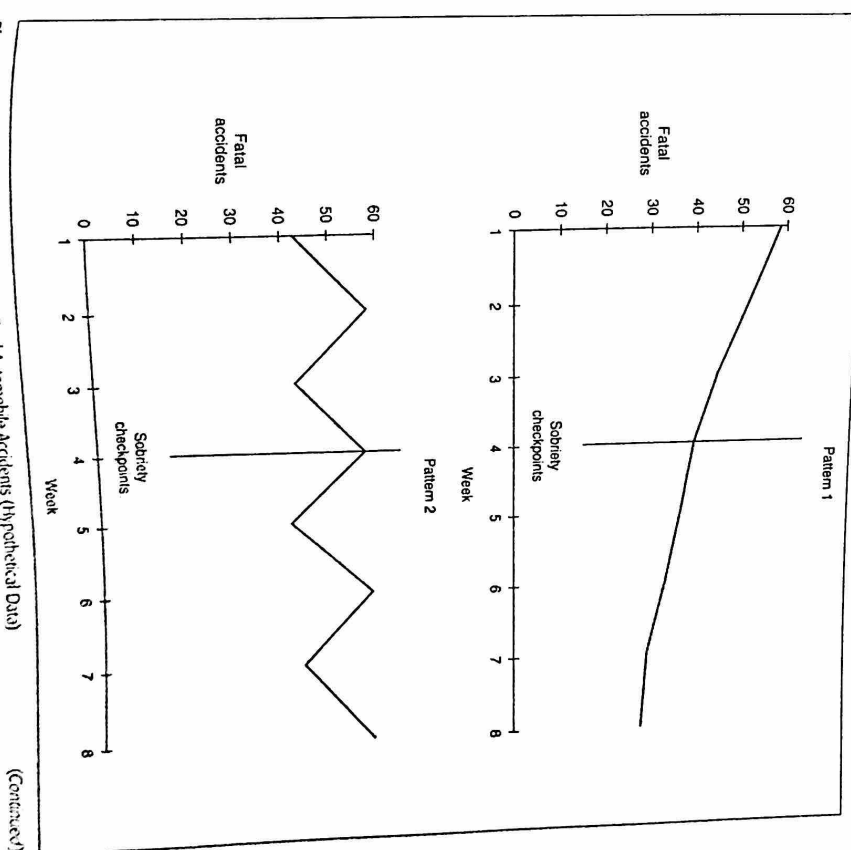
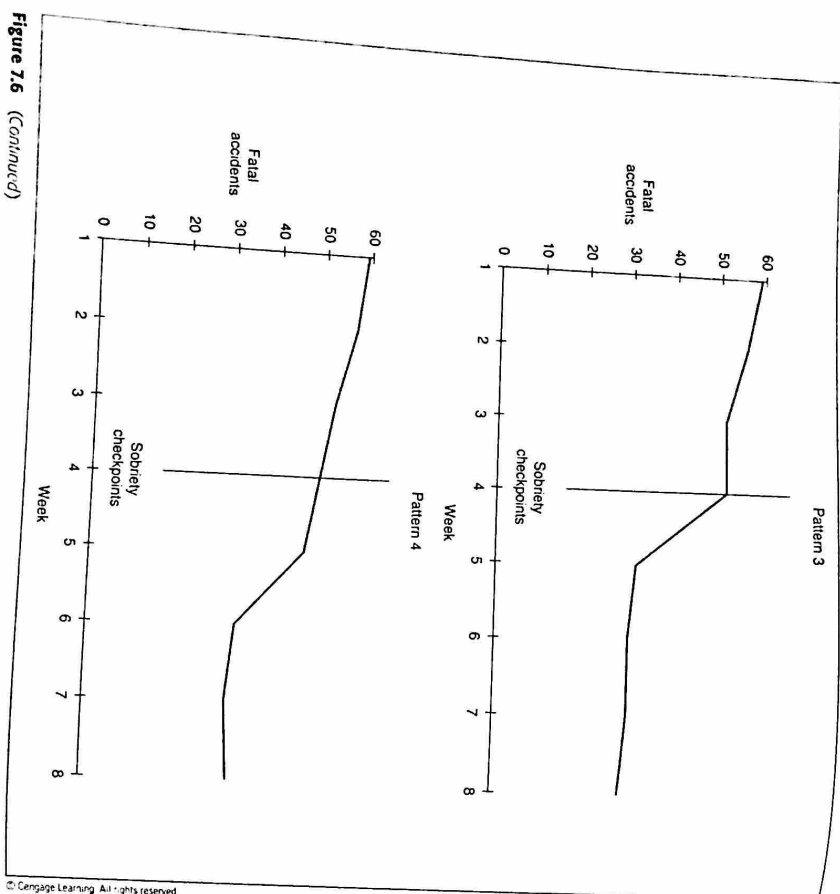


Figure 7.6 Four Patterns of Change in Fatal Automobile Accidents (Hypothetical Data)

(Continued)



© Cengage Learning. All rights reserved.

could reasonably be expected to reduce burglary. A special unit might eventually be able to reduce the number of burglaries, after investigating incidents over a period of time and making arrests. But it is highly unlikely that changing investigative procedures would have an immediate impact. This discrepancy prompted McCleary and associates to look more closely at the policy change and led to their conclusion that the apparent decline in burglaries was produced by changes in record-keeping

practices. No evidence existed of any decline in the actual number of burglaries.

This example illustrates our discussion of instrumentation earlier in this chapter. Changes in the way police counted burglaries produced what appeared to be a reduction in burglary. Instrumentation can be a particular problem in time-series designs for two reasons. First, observations are usually made over a relatively long time period, which increases the likelihood of changes

in measurement instruments. Second, time-series designs often use measures that are produced by an organization such as a police department, criminal court, probation office, or corrections department. There may be changes or irregularities in the way data are collected by these agencies that are not readily apparent to researchers and that are, in any case, not subject to their control.

As another example, David Finkelhor and Lisa Jones (2004) systematically examine data showing a nationwide 40 percent decline in child sexual abuse cases from 1990 to 2000. Because child abuse had become a prominent concern in the 1990s, the researchers were especially interested in determining whether any changes in recording practices in one or more states might account for the decline. They systematically considered such explanations as changing definitions or data collection procedures, more stringent screening of cases by investigators, less reporting of sexual abuse cases, and other possible sources of instrumentation. After comparing different sources of data for a variety of states, Finkelhor and Jones conclude that at least some portion of the decline in reported cases is due to a real decline in the number of child sexual abuse cases. Their careful report is an excellent example of thinking through plausible threats to instrumentation in a time series.

Variations in Time-Series Designs

If we view the basic interrupted time-series design as an adaptation of basic design building blocks, we can consider how modifications can help control for many validity problems. The simplest time-series design studies one group—the treatment group—over time. Rather than making one pretest and one posttest observation, the interrupted time-series design makes a longer series of observations before and after introducing an experimental treatment.

What if we considered the other building blocks of experimental design? Figure 7.7 presents the basic design and some variations using the familiar O, X, and I notation. In the basic design, shown at the top of Figure 7.7, many pretest and posttest observations are made on a single group that receives some treatment. We could strengthen

this design by adding a comparison series of observations on some group that does not receive the treatment. If, for example, roadside sobriety checkpoints were introduced all over the state of Ohio but were not used at all in Michigan, then we could compare auto accidents in Ohio (the treatment series) with auto accidents in Michigan (the comparison series). If checkpoints caused a reduction in alcohol-related accidents, we would expect to see a decline in Ohio following the intervention, but there should be no change or a lesser decline in Michigan over the same time period. The second part of Figure 7.7 shows this design—an interrupted time series with a nonequivalent comparison group. The two series are not equivalent because we did not randomly assign drivers to Ohio or Michigan. Wesley Skogan and associates present a good example of this design in their analysis of police problem solving in Chicago (Skogan et al., 2008). They examined changes in crime for police beats where specific problems were targeted compared with beats where no crime-specific interventions were developed.

A single-series design may be modified by introducing and then removing the intervention.

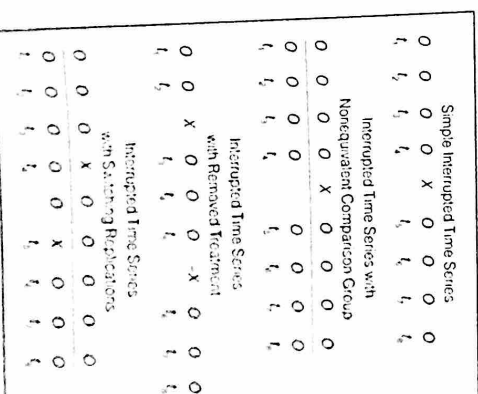


Figure 7.7 Interrupted Time-Series Designs

as shown in the third part of Figure 7.7. We might test sobriety checkpoints by setting them up every weekend for a month and then not setting them up for the next few months. If the checkpoints caused a reduction in alcohol-related accidents, we might expect an increase after they were removed. Or the effects of weekend checkpoints might persist even after we removed them.

Because different states or cities sometimes introduce new drunk-driving programs at different times, we might be able to use what Shadish, Cook, and Campbell (2002:192) called a time-series design with switching replications. The bottom of Figure 7.7 illustrates this design. For example, assume that Ohio begins using checkpoints in May 1998 and Michigan introduces them in July of the same year. A switching replications design could strengthen our conclusion that checkpoints reduce accidents if we saw that a decline in Ohio began in June and a similar pattern was found in Michigan beginning in August. The fact that similar changes occurred in the dependent variable in different states at different times, corresponding to when the program was introduced, would add to our confidence in stating that sobriety checkpoints actually reduced auto accidents.

Variable-Oriented Research and Scientific Realism

Another way to think about a time-series design is as a study of one or a few cases with many observations. If we design a time-series study of roadside checkpoints in Ohio, we will be examining one case (Ohio) with many observations of auto accidents. Or a design that compares Ohio and Michigan will examine many observations for two cases. Thinking once again about design building blocks, notice how we have slightly restated one of those building blocks. Instead of considering the number of experimental and control groups, our attention centers on the number of subjects or cases in our study. In Figure 7.7, the first and third time-series designs have one case each; the second and fourth designs examine two cases each.

Classical experiments and quasi-experiments with large numbers of subjects are examples of

what Charles Ragin (2000) terms **case-oriented research**, in which many cases are examined to understand a small number of variables. Time-series designs and case studies are examples of **variable-oriented research**, in which a large number of variables are studied for a small number of cases or subjects. Suppose we wish to study inmate-on-inmate assaults in correctional facilities. Using a case-oriented approach, we might send a questionnaire to a sample of 500 correctional facilities, asking facility staff to provide information about assaults, facility design, inmate characteristics, and housing conditions. Here, we are gathering information on a few variables from a large number of correctional facilities. Using a variable-oriented approach, we might visit one or a few facilities to conduct in-depth interviews with staff, observe the condition of facilities, and gather information from institutional records. Here, we are collecting information on a wide range of variables from a small number of institutions.

The **case study** design is an example of variable-oriented research. Here, the researcher's attention centers on an in-depth examination of one or a few cases on many dimensions. Robert Yin (2008) points out that the terms *case* and *case study* are used broadly. Cases can be individual people, neighborhoods, correctional facilities, courtrooms, or other aggregations. Our running series of case studies. Researchers gathered information from different sources to measure several types of variables for one state (New Jersey, North Carolina, Pennsylvania, etc.) and a varying number of other cases (state police stations, or counties) in the state.

Robert Yin cautions that the case study design is often misunderstood as representing "qualitative" research or a participant observation study. Instead, Yin advises that the case study is a design strategy and that the labels *qualitative* and *quantitative* are not useful ways to distinguish design strategies. Case studies might appear qualitative because they focus on one or a small number of units. But many case studies employ sophisticated statistical techniques to examine many variables for those units. An example illustrates how

misleading it can be to associate case studies with qualitative research.

In what has come to be known as the "Boston Gun Project," Anthony Braga and associates (2001) studied violence by youth gangs in Boston neighborhoods. Theirs was an applied explanatory study. They worked with local officials to better understand gang violence, develop ways to reduce it, and eventually assess the effects of their interventions. Neither a classical experiment nor a nonequivalent-groups design was possible. Researchers sought to understand and reduce violence by all gangs in the city. Their research centered on gangs, not individuals, although some interventions targeted particular gang members.

Researchers collected a large amount of information about gangs and gang violence from several sources. Earlier reports (Kennedy, Pichl, and Braga, 1996) described something called "network analysis," in which researchers examined relationships between gangs in different neighborhoods and conflicts over turf within neighborhoods. Police records of homicides, assaults, and shootings were studied. Based on extensive data on a small number of gangs, researchers collaborated with public officials, neighborhood organizations, and a coalition of religious leaders—the "faith community." A variety of interventions were devised, but most were crafted from a detailed understanding of the specific nature of gangs and gang violence as they existed in Boston neighborhoods.

David Kennedy (1998) summarizes these using the label "pulling levers," signifying that key gang members were vulnerable to intensive monitoring via probation or parole. The package of strategies was markedly successful: youth homicides were reduced from about 35–40 each year in the 20 years preceding the program to about 15 per year in the first 5 postintervention years (Braga, 2008).

The Boston research is also a good example of the scientific realist approach of Ray Pawson and Nick Tilley (1997). Researchers examined a small number of subjects—gangs and gang members—in a single city and in the context of specific neighborhoods where gangs were active. Extensive data were gathered on the mechanisms of gang violence. Interventions were tailored to those mechanisms in their context.

CHAPTER 7 Experimental and Quasi-Experimental Designs 193

Braga and associates (2001) emphasize that the success of the Boston efforts was due to the process by which researchers, public officials, and community members collaboratively studied gang violence and then developed appropriate policy actions based on their analysis. Other jurisdictions mistakenly tried to reproduce Boston's interventions, with limited or no success, failing to recognize that the interventions were developed specifically for Boston. In case study language, researchers examined many variables for one site and based policy decisions on that analysis. In the words of scientific realism, researchers studied the gang violence mechanism in the Boston context. In other contexts (Baltimore or Minneapolis, for example), gang violence operated as a different mechanism; the "levers" pulled in Boston did not work elsewhere. Braga and associates emphasize that the problem-solving process is exportable to other settings but that the interventions used in Boston are not (2001:220).

How do case studies address threats to validity? In the most general sense, case studies attempt to isolate causal mechanisms from possible confounding influences by studying very precisely defined subjects. Donald Campbell (2003:ix-x) likened this to laboratory experiments in the natural sciences, in which researchers try to isolate causal variables from outside influences. Case study research takes place in natural field settings, not in laboratories. But the logic of trying to isolate causal mechanisms by focusing on one or a few cases is a direct descendant of the rationale for experimental isolation in laboratories.

Figure 7.8 summarizes advice from Yin (2008:40–41) on how to judge the quality of case

Case Study Approach	
Construct Validity	Multiple sources of evidence Establish chain of causation
Internal Validity	Member checks Pattern-matching
External Validity	Time-series analysis Replicate through multiple case studies

Figure 7.8 Case Studies and Validity
Source: Adapted from Yin (2008:41).



EVALUATING EXPLANATIONS FOR DISPROPORTIONALITY IN TRAFFIC STOPS

The kind of research we have been describing with respect to racial profiling is generally explanatory research. In a sense, we want to learn whether race is associated with differences in police action. Recalling the time-order criterion for demonstrating cause, because an individual's race is determined before any particular experience with traffic enforcement, differences in traffic enforcement by race may be explained by race. We emphasize *may*, because time order is only one of the three requirements for demonstrating cause.

Experiments in Racial Profiling?

You may have heard something about field experiments to assess differences between the experiences of black and white people applying for a mortgage, shopping for a new car, renting an apartment, or buying a house. Individuals, or couples in the case of home buying, are matched on age, income, and other criteria relevant to whatever transaction is being tested. Race is systematically varied acting as an experimental stimulus. The goal is to assess any discriminatory treatment where the only difference between people is race (National Research Council, 2004).

It's not possible to systematically vary race in this kind of experimental fashion and test for different outcomes in criminal justice processing. We can't control or systematically vary the independent

variable in the general sense of the word "experiment" as described by Farrington et al. (1986). Imagine, for example, arranging for teams of researchers, half white and half African American, driving identical vehicles at 20 mph above the posted speed limit on an interstate highway. It's interesting to think about doing that kind of research but you will quickly recognize that it can't legally or safely be done. Remember ethics from Chapter 3?

Quasi-Experiments?

Research on race and traffic enforcement is either quasi-experimental or variable-oriented research such as case studies. Because we have been describing research primarily in three states, it might seem that most such efforts are case studies. However, the strongest research—in North Carolina and Pennsylvania—uses a sophisticated series of comparisons to sort out what causes race disparities in traffic stops. These have come to be called benchmarks.

Researchers try to compare a "treatment group," minorities as a proportion of traffic stops, to some benchmark, or comparison group. As Engel et al. (2004:8) put it, the comparison group is intended to reflect: "... the 'expected' rate of stops of minorities assuming that no racial discrimination exists." Here are some examples of comparison groups, together with comments on their strengths and weaknesses.

- Resident population. If nonwhites are 25 percent of drivers stopped, but are only 15 percent of the resident population of a jurisdiction, this suggests a racial disparity. The problem with this comparison is that drivers are mobile by definition. The majority of cars on the southern third of the New Jersey Turnpike have out-of-state license plates. This approach might work on local streets in a large jurisdiction, but even then it ignores the fundamental fact that people are mobile. Researchers in England recognized this in their study of four municipalities (Miller, 2000).
 - Driving population. This is better because it compares people stopped by police to people who are actually on the road. In settings like the New Jersey Turnpike, the race of drivers can be observed at toll-booths, as we described in Chapter 5. However, this method produces information about drivers, not traffic violators. Police claim to stop cars because they observe some violation, not because of race.
 - Night-time traffic stops. This comparison strategy assumes that police are not able to detect the race of drivers at nighttime. So if the proportion of traffic stops of nonwhite drivers at night is lower than the proportion stopped during the day, some discriminatory targeting might be at work in daylight hours. Even though there is merit in such logic, New Jersey state police tend to make fewer stops at night. There have also been reports of "spotlighting" by state troopers, where they use spotlights, or park their vehicle at right angles to the roadway and shine lights on the face of passing drivers (Venierio and Zoubek, 1999).
 - Radar-timed stops. Similar logic is at work here. If police make stops based on radar readings, it's assumed they will be less likely to stop people based on the appearance of race. In this respect, police are presumed to exercise less discretion in relying solely on radar to decide whom to stop. Again, it's not quite that simple. From their research in New Jersey, Maxfield and Andersen sat in state police cars while troopers used radar. In many cases the racial appearance of drivers was easily visible. Most vehicles were speeding, though some more than others. Together, these observations suggest that state police could use race as a factor in deciding which speeding cars to stop and which to ignore.
 - Independent observations of speeding. As we saw in the last chapter, researchers in three states collected their own data on speeding. In principle this offers the best comparison against police stop data: comparing the race distribution of police stop data to the race distribution of speeders from data collected independently. If the percentages are similar, it supports police claims that they stop only traffic violators. If they are different, this would be stronger evidence of unexplained disparity.
- We say "unexplained disparity" intentionally. Such a disparity would support discrimination as an explanation, but it would not prove that differences were due to discrimination. That's because we cannot be as certain as we would like that all other plausible explanations have been eliminated. Slightly restated, we cannot rule out threats to internal validity of a claim that discrimination causes disparities.

study designs in language that should now be familiar. Construct validity is established through multiple sources of evidence, the establishment of chains of causation that connect independent and dependent variables, and what are termed "member checks"—asking key informants to review tentative conclusions about causation. Examples of techniques for strengthening internal validity are theory-based pattern matching and time-series analysis. The first criterion follows Shadish, Cook,

and Campbell, calling on researchers to make specific theory-based predictions about what pattern of results will support hypothesized causal relationships. Alternative explanations, also termed "rival hypotheses," are less persuasive when specific predictions of results are actually obtained. For example, Braga and associates (2001) predicted that gun killings among male Boston residents under age 25 would decline following implementation of the package of interventions in the Boston

gun strategy. Although other explanations are possible for the sharp observed declines, the specific focus of the researchers' interventions and the concomitant results undermine the credibility of rival hypotheses. Having many measures of variables over time strengthens internal validity if observations support our predicted expectations about cause. We saw earlier how nonequivalent time-series comparisons and switching replications can enhance findings. This is also consistent with

pattern matching—we make specific statements about what patterns of results we expect in our observations over time.

Finally, a single case study is vulnerable to external validity threats because it is rooted in the context of a specific site. Conducting multiple case studies in different sites illustrates the principle of replication. By replicating research findings, we accumulate evidence. We may also find that causal relationships are different in different

settings, as did researchers who tried to transplant specific interventions from the Boston Gun Project. Although such findings can undermine the generalizability of causality, they also help us understand how causal mechanisms can operate differently in different settings.

Time-series designs and case studies are examples of variable-oriented research. A case study with many observations over time can be an example of a time-series design. Adding one or more other cases offers opportunities to create nonequivalent comparisons. Time-series designs, case studies, and nonequivalent comparisons are quasi-experimental designs—they are conducted in the manner of experiments, using design building blocks in different ways.

Experimental and Quasi-Experimental Designs Summarized

Understanding the building blocks of research design and adapting them accordingly works better than trying to apply the same design to all research questions.

By now, it should be clear that there are no simple formulas or recipes for designing an experimental or quasi-experimental study. Researchers have an almost infinite variety of ways of varying the number and composition of groups of subjects, selecting subjects, determining how many observations to make, and deciding what types of experimental stimuli to introduce or study. See the discussion of different approaches in our running example, “Putting It All Together: Evaluating Explanations for Disproportionality in Traffic Stops.”

Variations on the classical experimental designs are especially useful for exploratory research and in evaluation studies, but exploratory and descriptive studies usually use other methods. Surveys conducted at one point in time, for example, may be used to explore or describe such phenomena as fear of crime or public attitudes toward punishment. Longitudinal studies of age cohorts

are often the best way to examine criminal careers or developmental causes of delinquency.

Even when experimental designs might be the best choice, it is not always possible to construct treatment and control groups, to use random assignment, or even to analyze a series of observations over time. Research on CCTV is an example.

As we stated early in this chapter, experiments are best suited to topics that involve well-defined concepts and propositions. Experiments and quasi-experiments also require that researchers be able to exercise, or at least approximate, some degree of control over an experimental stimulus. Finally, these designs depend on the ability to unambiguously establish the time order of experimental treatments and observations on the dependent variable. Often, it is not possible to achieve the necessary degree of control.

In designing research projects, researchers should be alert to opportunities for using experimental designs. Researchers should also be aware of how quasi-experimental designs can be developed when random assignment is not possible. Experiments and quasi-experiments lend themselves to a logical rigor that is often much more difficult to achieve in other modes of observation. The building blocks of research design can be used in creative ways to address a variety of criminal justice research questions. Careful attention to design issues, and to how design elements can reduce validity threats, is essential to the research process.

SUMMARY

- Experiments are an excellent vehicle for the controlled testing of causal processes. Experiments may also be appropriate for evaluation studies.
- The classical experiment tests the effect of an experimental stimulus on some dependent variable through the pretesting and posttesting of experimental and control groups.
- It is less important that a group of experimental subjects be representative of some larger population than that experimental and control groups be similar to each other.
- Random assignment is the best way to achieve comparability in the experimental and control groups.

CHAPTER 7 Experimental and Quasi-Experimental Designs 197

- The classical experiment with random assignment of subjects guards against most of the threats to internal validity.
- Because experiments often take place under controlled conditions, results may not be generalizable to real-world situations, or findings from an experiment in one setting may not apply to other settings.
- The classical experiment may be modified to suit specific research purposes by changing the number of experimental and control groups, the number and types of experimental stimuli, and the number of pretest or posttest measurements.
- Quasi-experiments may be conducted when it is not possible or desirable to use an experimental design.
- Nonequivalent-groups and time-series designs are two general types of quasi-experiments.
- Time-series designs and case studies are examples of variable-oriented research, in which a large number of variables are examined for one or a few cases.
- Both experiments and quasi-experiments may be customized by using design building blocks to suit particular research purposes.
- Not all research purposes and questions are amenable to experimental or quasi-experimental designs because researchers may not be able to exercise the required degree of control.

KEY TERMS

- Case-oriented research (p. 192)
- Case study (p. 192)
- Classical experiment (p. 169)
- Control group (p. 170)
- Dependent variable (p. 169)

REVIEW QUESTIONS AND EXERCISES

1. If you do not remember participating in DARE—Drug Abuse Resistance Education—you have probably heard or read something about it. Describe an experimental design to test the causal hypothesis that DARE reduces drug use. Is your experimental design feasible? Why or why not?
2. Experiments are often conducted in public health research, where a distinction is made between an efficacy experiment and an effectiveness experiment. Efficacy experiments focus on whether some new health program works under ideal conditions; effectiveness experiments test the program under typical conditions that health professionals encounter in their day-to-day work. Discuss how efficacy experiments and effectiveness experiments reflect concerns about internal validity threats on the one hand and generalizability on the other.
3. Crime hot spots are areas where crime reports, calls for police service, or other measures of crime are especially common. Police in departments with a good analytic capability routinely identify hot spots and launch special tactics to reduce crime in these areas. What kinds of validity threats should researchers be especially attentive to in studying the effects of police interventions on hot spots?