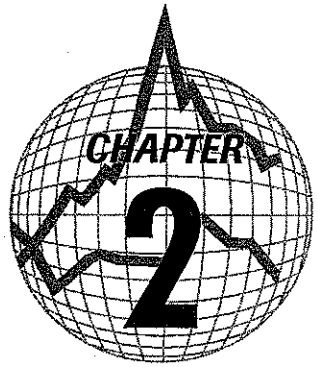


Research Design

CHAPTER OBJECTIVES

After reading this chapter, you should be able to

- Distinguish between independent and dependent variables
- Describe the six steps of program evaluation
- Explain experimental and quasi-experimental designs
- Understand the importance of rival hypotheses
- Identify threats to validity in research design

This chapter introduces major concepts in social science research and applies them to program evaluation. Program evaluation, which helps managers and analysts to determine the outcomes of programs and policies, is an important and necessary skill for managers and analysts to have. This chapter also examines a variety of research designs commonly used in program evaluation.

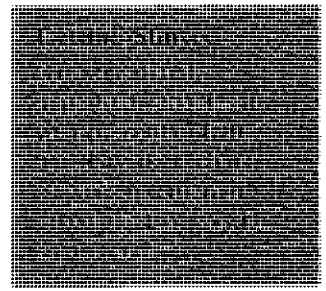
INTRODUCING VARIABLES AND THEIR RELATIONSHIPS

Research is fundamentally about establishing the nature of things. For example, assume that we are responsible for managing a program to reduce

high school violence or that we are otherwise interested in this topic. One of the first steps that we need to take is to gain a solid understanding of this phenomenon, high school violence, by examining the ways in which it is manifested. We would want to know about its various forms such as verbal and emotional abuse; its physical manifestations such as shoving, hitting, and the use of weapons; and its racial and sexual manifestations, too. Thereafter, we would want to know the magnitude of each manifestation, such as how many fist fights, gun fights, or rapes occur. And we might want statistics on specific types of injuries, such as broken bones or concussions. Indeed, the frequency of these phenomena often is a key target for management and public policy.

The same is true for many other phenomena such as community conditions (for example, poverty or economic growth), events (such as wildfires or toxic spills), as well as programs and policies shaping conditions and events. We will want to first establish the manifestations of a phenomenon and then learn something about their magnitude. If we are interested in environmental quality, for example, we will want to know facts about the state of the environment and how it varies in different ways and in different locations. If we are in health care management, we will want to know the incidence of different diseases. If we are interested in inflation, we will want to know its current level, which factors such as energy prices or housing costs are responsible for recent changes, and how inflation varies in different parts of the country. Once we decide what we are interested in, we will want to know more about its manifestations and variations.

Public and nonprofit management and policy typically involve phenomena that vary in some way. ***Variables*** are defined as empirically observable phenomena that vary. This is best illustrated by a few examples. "High school violence" is a variable because it is observable and varies across schools; violence is more common in some schools than in others, and we can observe the differences. "Environmental quality" is also a variable because it is observable and varies across locales, as do "diseases" and "inflation," for example. Variables are key to research, and they are everywhere. The number of students in classes is also a variable because different classes have different numbers of students, and the number of students in each class can be observed. By contrast, in a study of only female students, the variable "gender" does not vary and is therefore called a ***constant.*** Constants are phenomena that do not vary.



Attributes are defined as the specific characteristics of a variable, that is, the specific ways in which a variable can vary. All variables have attributes.

For example, high school violence can be measured as being absent, sporadic, occurring from time to time, or ongoing—these are the attributes of the variable "high school violence." Another example is the variable "gender." Gender varies in the population, and the attributes of gender are "male" and "female." The variable "race" often has more than two attributes (Caucasian, African American, Native American, and so forth). The variable "income" can have few or a nearly infinite number of attributes if income is measured as specific dollar amounts. In surveys, often each survey item is treated as a separate variable, and the response categories for each question are the variable's attributes. For example, the question "What is your gender?" is considered a variable, and the response categories "male" and "female" are its attributes.

Research usually involves both *descriptive analysis* and the study of *relationships* involving variables. ***Descriptive analysis*** provides information about the nature of variables—such as whether a high school violence problem exists and the extent or level of it. The preceding section gave examples of descriptive analysis. In our high school violence example, descriptive analysis can be used to show the nature of the perpetrators, the geographic areas in which such violence most often occurs, and the extent to which it is perceived as a problem. Descriptive analysis is useful in public management and policy because managers need to know the state of the world that they are trying to shape. They need to know, for example, the number of teenagers who have been hurt by others at school. This is simply a number—such as 5 percent.

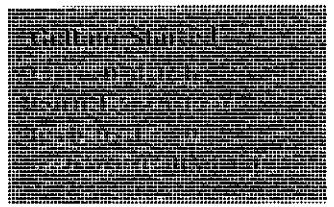
Managers also want to know the causes of problems and the effectiveness of interventions. This involves examining ***relationships***, that is, specifying which variables are related to each other, and the ways in which they are related. Indeed, research is not only about establishing the nature of phenomena, but also about their relationships. For example, we might want to examine whether students who participate in anger management classes describe themselves as being less angry or less prone to acting out against others. Specifically, we want to know whether participation in anger management class decreases the extent of acting out by students. We might also examine the effect of other conditions—such as drug use or gang participation—on high school violence. By knowing how programs and conditions affect outcomes, managers can better recommend and pursue alternative courses of action. Most studies involve both descriptive analysis and an examination of relationships.

Relationships in social science are distinguished by whether they are *probabilistic* (occurring sometimes) or ***deterministic*** (occurring each time). For example, when we say that anger management reduces high school violence, we are not implying that this always occurs, for each student. Some

students might even become more violent, perhaps learning new ways of expressing their anger. Rather, we mean that, *on average*, the number of violent incidents will decrease. The number of incidents will decrease for some students more than for others, and for still others it will not decrease at all; the relationship is probabilistic in nature. Many relationships in the social world are probabilistic.

When social scientists say that "anger management reduces high school violence," they typically mean that in most instances anger management reduces high school violence. They usually also have a standard in mind, such as anger management reducing high school violence at least 95 out of 100 times. Sometimes, social scientists adopt an even stricter standard, such as at least 99 out of 100 times. By adopting such standards, social scientists provide information about probabilistic relationships with a relatively high degree of confidence.[1]

Relationships also are distinguished as being either *causal* or *associational*. ***Causal relationships*** show cause and effect, such as the impact of anger management programs on high school violence, the impact of employee compensation on workplace productivity, or the impact of environmental policies on water quality. In these instances, one variable is assumed to affect another. By contrast, associations are relationships that imply no cause and effect. For example, it is said that in Sweden a relationship exists between the number of storks and the number of childbirths; both increase in the spring. Does this imply that storks really do bring babies, at least in Sweden? No, of course it doesn't. The appearances of storks and new babies are unrelated; they have no cause-and-effect relationship.[2]



Among causal relationships, we further distinguish between *independent variables* and *dependent variables*. ***Dependent variables*** are variables that are affected by other variables (hence, they are dependent on them). ***Independent variables*** are variables that cause an effect on other variables but are not themselves shaped by other variables (hence, they are independent). For example, in a study of the impact of anger management on high school violence, anger management is the independent variable that affects high school violence, which is the dependent variable. Causal relationships are commonly thought of in the following manner:

Independent Variable(s) → Dependent Variable

An important step in any research is specifying the dependent and independent variables. Doing so brings clarity and direction to the research.

Although many studies examine several relationships, most evaluations focus on explaining only a few dependent variables. In our example, we wish to examine the impact of anger management on high school violence:

Independent Variable		*Dependent Variable*
Anger Management	→	High School Violence

Of course, our evaluation needn't be limited to studying just this relationship, but specifying relationships in this manner helps concentrate our attention on (1) accurately determining the level of high school violence and (2) examining whether anger management is associated with it. We might also study the effect of gun control laws (independent variable) on this dependent variable, or other relationships such as the effect of homework assistance (independent variable) on academic performance (dependent variable). *Distinguishing between independent and dependent variables clarifies and sharpens one's thinking about which variables are being studied and how they are related to each other. It is a cornerstone of research, program evaluation, and policy analysis, and it is an essential skill that managers and analysts will want to practice.*

A literature review of scholarly (research) and professional articles can often help to further develop and clarify our thinking about independent and dependent variables. Oftentimes, managers and analysts are interested in a phenomenon, such as school violence, and perhaps one or two factors associated with it. Then, prior research can help to further develop this interest. Previous studies may suggest ways of measuring high school violence and perhaps provide a critical review of alternative measures. Researchers might have also taken different perspectives on the causes of high school violence, leading them to consider different independent variables. Research might have evaluated the effect of independent variables in different settings, though not necessarily yours. Research might have carried this interest further, examining the impact of high school violence on, for example, educational performance. In these different ways, prior research, as a reflection of careful and considered thought, can be used by managers and analysts in helping them to further develop their topic.

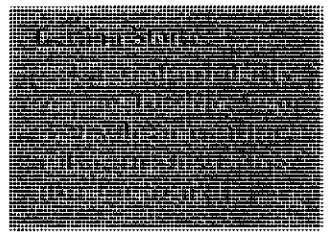
Program evaluation is often intended to stake a claim of *causation*. In our example, managers might want to argue that anger management has caused the decline in high school violence. You may have heard the expression "correlation does not prove causation." This is true. Causation requires both (1) *empirical (that is, statistical) correlation* and (2) *a plausible cause-and-effect argument*. These two ***criteria for causality*** must be present. Statistical analysis tests whether two variables are correlated, but causality also

requires a persuasive argument (also called theory) about how one variable could directly affect another.[3] Regarding the impact of anger management on high school violence, a plausible theory might readily be written up. Anger management training teaches people how to identify anger and release it in ways that are nonviolent toward others. Thus, both statistical correlation and a persuasive theoretical argument are required to stake a claim of causation.

How difficult can it be to make a theoretical argument of cause and effect? Examining, say, the relationship between gender and high school violence, we have yet to make a plausible cause-and-effect argument. If we lack specific evidence (especially evidence that might persuade a skeptical audience) that gender, defined by reproductive organs and hormones, causes violence, then we best regard this relationship as a mere correlation, that is, an ***association***. Empirical correlations remain mere associations until analysts have argued, in persuasive and exacting detail, how one variable can plausibly cause another.

Finally, relationships that have not yet been empirically tested (that is, established) are called ***hypotheses***. For example, a study hypothesis might be that, on average, female teenagers are less prone to violence than males. Then, empirical data will need to be collected and analyzed in order to prove the hypothesis either true or false for the population from which these data are drawn. Subsequent chapters in this book discuss how to analyze data and draw conclusions about hypotheses. Academic research studies are usually quite explicit about which hypotheses are being tested and why they are relevant.

This brief introduction lays out important concepts that are used over and over again in research. Quite simply, when we do research, we see the world existing of variables and their relationships. We also identify the attributes of variables, and ask whether relationships are deterministic or probabilistic, causal or only associational, tested (established) or hypothesized.



PROGRAM EVALUATION

Program evaluation can be defined as the use of social science research methods to determine whether, and in what ways, a program works. Program evaluation involves the description of programs, conditions, and events, as well as the analysis of relationships, such as the impact of programs on outcomes. Program evaluation uses both quantitative and qualitative methods to describe programs and analyze their relationships.

How difficult can it be to document program outcomes? There usually is more to program evaluations than meets the eye. Among the first challenges is to find out what the program is expected to accomplish. Consider the following example. In response to growing concerns about teen violence, many communities and states have created after-school programs. The idea, according to elected officials and supported by the public, is to get teenagers off the streets and into supervised environments. As a public manager, your job is to implement such a program. Funding guidelines require that you document the success of the program.

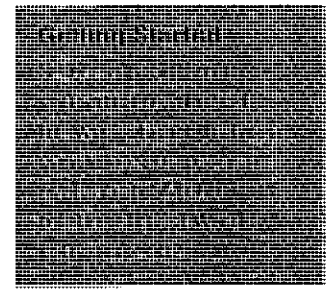
Now, you must figure out what the program is expected to accomplish. You might be surprised to learn that sometimes little thought has gone into identifying specific outcomes for such programs, or that some elected officials and experts have different views. Some advocates only want teenagers off the streets, but others expect them to learn something as well. Still others feel that anger management should be taught. Even if you are responsible only for program evaluation, oftentimes you will find yourself formulating program outcomes.

Next, assume that you and others agree that anger management is one of several appropriate activities for the after-school program. Specifically, the after-school program will teach students to recognize and deal with anger in appropriate ways. You might even try to target so-called high-risk students. How will you measure the success of your anger management efforts? Should you ask students whether they feel less angry? Should you ask their parents and teachers as well? Should you ask teachers to record the number of classroom incidents, such as student outbursts? Should you do all of this? If so, in what way?

Suppose you decide to ask teachers to track classroom incidents. Which incidents should be tracked? Is it appropriate to compare different classroom incidents across schools or classes? Should you develop baseline data, and if so, which? Also, how accurate do you think the teachers will be in their reporting and tracking? Are their responses likely to be biased in any way? Or suppose you decide to send a survey to parents. Do you need to send your survey to all parents? How many questions should you ask? What response rate is appropriate? How do you avoid biased questions?

Finally, consider the possibility that the number of classroom incidents drops during the course of your anger management program. How do you know that the drop is due to the anger management course? Could teachers and parents have become more involved in anger management themselves? What if some students who are known to be angry and violent were transferred out of the school? In short, how sure can you be that any changes are due to the after-school program?

These questions are hardly academic. Elected officials and senior managers expect others to have answers to such questions, regardless of whether they concern after-school programs, prison overcrowding, environmental protection, or national security. Determining which outcomes ought to be measured and measuring their attainment in credible ways are activities germane to all public programs and polices. Public departments need people with skills to assess program outcomes; program evaluation applies social science methods to these issues.[4]



Six Steps

Program evaluation usually involves six steps. The purpose of these steps is to help researchers and managers identify and address relevant concerns in an orderly manner. These steps help ensure that evaluation is done in objective and scientifically valid ways—evaluation findings must be credible and stand up under the light of public scrutiny—and that conclusions and recommendations are embraced by those who have the power to bring about change. Program evaluation must include opportunities for stakeholders to have input; study conclusions must be credible, relevant, and consistent with opportunities for change. The following ***six steps of program evaluation*** provide a strategic road map that combines these dual needs—to be both responsive and objective:

1. *Define the activity and goals that are to be evaluated.* What are the key objectives and constraints according to key decision makers? What are the main objectives and concerns according to program staff? How do clients and others outside the program view it? What is the key target population of these activities and goals?
2. *Identify which key relationships will be studied.* Which program outcomes does the evaluation measure? Which factors are hypothesized to affect these program outcomes? Which counter-explanations are considered?
3. *Determine the research design that will be used.* Will a control or comparison group be used? Is there a need for developing a baseline of current performance? Are periodic or follow-up measurements foreseen and, if so, over what time period?
4. *Define and measure study concepts.* Which study concepts require detail in measurement? Which concepts require little detail? Will existing data be used, and how accurate are they? Will new data be gathered through, for example, a survey or focus group? If so, who will undertake such a

project, and how long will it take? What statistical requirements must the data meet for subsequent analysis? What resources and expertise are needed for data collection and program evaluation? What suggestions do key decision makers and others have for improving measurement?

5. *Collect and analyze the data.* Which statistical techniques will be used for data analysis? What type of conclusions are researchers seeking from the data? Do the data meet the requirements of different statistical techniques?
6. *Present study findings.* How, and to whom, will conclusions be presented? Can presentations be part of other consensus and decision-making processes? Can preliminary feedback about tentative findings be obtained from key decision makers and others? Who requires a detailed analysis and presentation? Who requires only a brief overview of main findings? What should the final report look like, and to whom should it be sent?

Previously we dealt with some matters pertaining to the first two steps. In our example, the activity is anger management as an after-school program. This program will teach students to recognize and deal with their anger. At this point, we might further specify that the program targets high-risk students, though it may include other students as well. The preceding questions prompt us to make finer specification. Also, assume that after further interviewing school administrators, teachers, and students and their parents, additional program objectives are formulated, in addition to reducing high school violence. These additional objectives might be to keep students safe after school, to improve academic performance, to reduce disruptive classroom behavior, to reduce violent behavior outside school, and to provide opportunities for getting involved in other, "fun" activities such as sports and music. The latter might seem far removed from anger management objectives, but program clients sometimes view such activities as useful motivators for continued participation.

Regarding the second step, while participation in anger management is examined for its impact on the specified outcomes, it is recognized that other factors might play a role, too. For example, gang participation and drug use is likely to reduce the effect of anger management training as a result of strong countervailing peer pressures and addictive impulses. Also, a lack of parental interest in their children's education is a likely negative factor. On the other hand, being transferred to a low violence school might reduce violence. Thns, program evaluation will need to consider additional circumstances along with the impact of anger management training. Such circumstances are part of steps 2 and 3 and are discussed below. Step 4, the definition and measurement of study concepts, is discussed in Chapter 3.

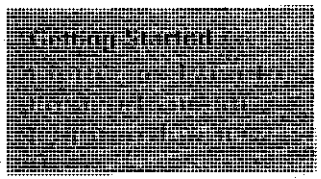
Rival Hypotheses and Limitations of Experimental Study Designs

The purpose of research design is to help ascertain that outcomes, such as reduced high school violence, are occurring and plausibly related to the program and not to other factors. But what if, parallel to anger management, another program aims to reduce student access to weapons? Then it is conceivable that any reduction in school violence might be partly or entirely ascribed to this other program. Such alternative explanations for observed outcomes are called ***rival hypotheses***, and variables used to measure rival hypotheses are called ***control variables***. Control variables are empirical, just as dependent and independent variables are, but they get their name from their research role: to test whether relationships between independent and dependent variables hold up under the presence of alternative, rival explanations for the observed pattern of outcomes. They are sometimes also called confounding variables, referring to concomitant activities that also explain outcomes and, hence, complicate efforts to establish a causal effect of programs or policies on outcomes. In our example, the presence of a weapons access policy, a concomitant event, is certainly a control variable that the manager will want to take into account. Indeed, the credibility of research findings often rests on the extent to which pertinent rival hypotheses have been identified and incorporated into study designs.[5]

Rival hypotheses (and their associated control variables) can be dealt with through experimental design and statistical analysis. ***Experimental designs*** address rival hypotheses through the use of control groups, which are similar to the study group in all aspects *except* that members of the control group do not participate in the intervention. You may be familiar with control groups through literature that describes the effectiveness of medical treatments. In ***classic, randomized experiments***, participants are randomly assigned to either a control or an experimental (or study) group. The assignments are random to ensure that any observed differences between these two groups are due only to the treatment and not to any other factor. Random assignment ensures that the two groups are similar, and baseline data are used to further rule out any chance differences in the groups' respective starting conditions. Further, neither the participants of the control and study groups nor their doctors are told whether they are receiving the experimental treatment or the ineffective placebo (they both look alike), because doing so might cause patients or their doctors to alter their behavior. In short, everything is done to ensure that the *only* difference between the groups is that one gets the treatment and the other does not. The logical inference, then, is that any difference *must* be due to the experimental treatment. The research design rules out every other factor.

Programs and policies are the public management equivalent of clinical interventions. Unfortunately, classic, randomized experiments are notoriously difficult to implement in public administration and policy because it is generally legally and ethically impossible to deny citizens or jurisdictions programs and policies. In our example, we do not envision randomly assigning teenagers to after-school programs. Some parents would be outraged if their children were denied access to the anger management program. They might even sue. It is also unclear what the "placebo" intervention might be in our example; it is absurd to suggest that subjects might participate in an anger management program that is intentionally designed to be ineffective. The problems of rival hypotheses are real, but the classic, experimental design is seldom a feasible strategy for addressing this matter in public administration and public policy.

The fact that we are unable to conduct classic, randomized experiments in public programs does not mean that we cannot use comparison groups or baselines measurement. Indeed, doing so can add valuable information to our program evaluation. For example, it would be interesting to compare high school violence among schools, of which only some have anger management programs. The term *comparison group* rather than control group would then be used, because the comparison group is not similar in all ways to the experimental group but for the intervention; other differences may exist. Clearly we can no longer rely on the research design itself to rule out the presence of rival hypotheses; rather, we must use the ***strategy of statistical control*** to account for rival hypotheses. This strategy involves (1) identifying plausible rival hypotheses, (2) collecting data about them, and (3) using *statistical techniques* to examine their impact on high school violence, relative to anger management. Specifically, we ask, What is the impact of anger management on high school violence, controlled for these other factors? The statistical techniques for analyzing data in this way are discussed later in this book; however, this approach obviously requires that analysts identify relevant control variables and collect data about them prior to analysis. Hence, the need to identify relevant rival hypotheses is determined early in program evaluation, during step 3.



QUASI-EXPERIMENTAL DESIGNS IN PROGRAM EVALUATION

Comparisons between experimental and comparison groups that do not meet the standard of classic research designs are called ***quasi-experimental designs.*** These designs may lack randomization, baseline (or pretest) measurement, or a comparison group. However, comparison groups and baselines often provide important information that help evaluate the effectiveness of programs and policies. Comparison groups provide a useful reference, for example, when outcomes show a widening gap between groups. Without a baseline, it is harder to persuade others that a program has had an impact. These features should be considered as part of the research study whenever practicable.

It is useful to view quasi-experimental research designs as variations on the classic, randomized design. Box 2.1 provides a stylistic representation of such designs. Design A is the classic, randomized design, showing randomization, a control group, pretests, and posttests. The designs under B are all lacking in one or more ways. Specifically, they all lack randomization, and many also lack a pretest, a comparison group, or both. These are quasi-experimental designs. Designs B2, B3, and B4 are rather typical quasi-experimental designs in public and nonprofit management.

Sometimes policy analysts and managers want to assess the impact of a policy or program after it has been implemented, without taking steps to develop such an assessment prior to the intervention (design B2). Perhaps managers had not given prior consideration to conducting the outcome evaluation. Lacking systematic assessment prior to the intervention, after-the-fact designs often are limited to interviewing managers and participants about their subjective assessment of impacts, and gathering quantitative data that might provide insights about outcomes. Data from before the intervention may be available, such as through administrative records that are routinely gathered. However, such data vary in their pertinence to the intervention.

Sometimes a comparable comparison group is found (design B3). The art is to find such a group, such as a comparable school that does not have an anger management program. Analysts will have to argue that the comparison group is indeed a valid comparison group. In our example, they might find a school where students have similar test scores, similar socioeconomic backgrounds, similar academic and extracurricular programs, and similar arrest and felony rates among students. Then these data must be gathered and these variables used as control variables to statistically control for any differences that might exist. An example of such a design is discussed in Box 2.2.

In other instances, evaluation is planned prior to intervention. Sometimes evaluation of programs and policies is a requirement, which may spur such advance planning. Then, to supplement administrative data, additional baseline data are gathered, as well as information pertaining to rival hypotheses that might be considered (design B3). The design of before and after measures may vary. For instance, measurements may span several time periods, before and after the intervention. In our example, high school violence might be recorded on a monthly basis, 6 months prior to and perhaps up to 12 months after the intervention, producing time series data.

In Greater Depth...

Box 2.1 Research Designs

Research designs can be characterized using the following notation, where R = randomization, X = intervention, and O = measurement. The following is based on the enduring, classic work of Donald Campbell and Julian Stanley.

A. The classic, randomized design is depicted graphically as follows. Any significant program impact would be indicated when (O2–O1) > (O4–O3). The placebo intervention is not shown, but if it existed it would be implemented between O3 and O4; it would be similar to X, except that it is intentionally ineffective.

	Pretest	Program	Posttest
Group 1:	R O1	X	O2
Group 2:	R O3		O4

B. Quasi-experimental designs vary from this design in several ways:

1. Research design with a nonrandomized comparison group:

	Pretest	Program	Posttest
Group 1:	O1	X	O2
Group 2:	O3		O4

2. One-group research design with posttest measure, only:

	Pretest	Program	Posttest
Group 1, only:		X	O2

3. Research design with comparison group and posttests, only:

	Pretest	Program	Posttest
Group 1:		X	O2
Group 2:			O4

4. One-group research design with pretest and posttest:

	Pretest	Program	Posttest
Group 1, only:	O1	X	O2

Source: Donald Campbell and Julian Stanley, *Experimental and Quasi-experimental Designs for Research* (Chicago: Rand McNally), 1963.

In Greater Depth...

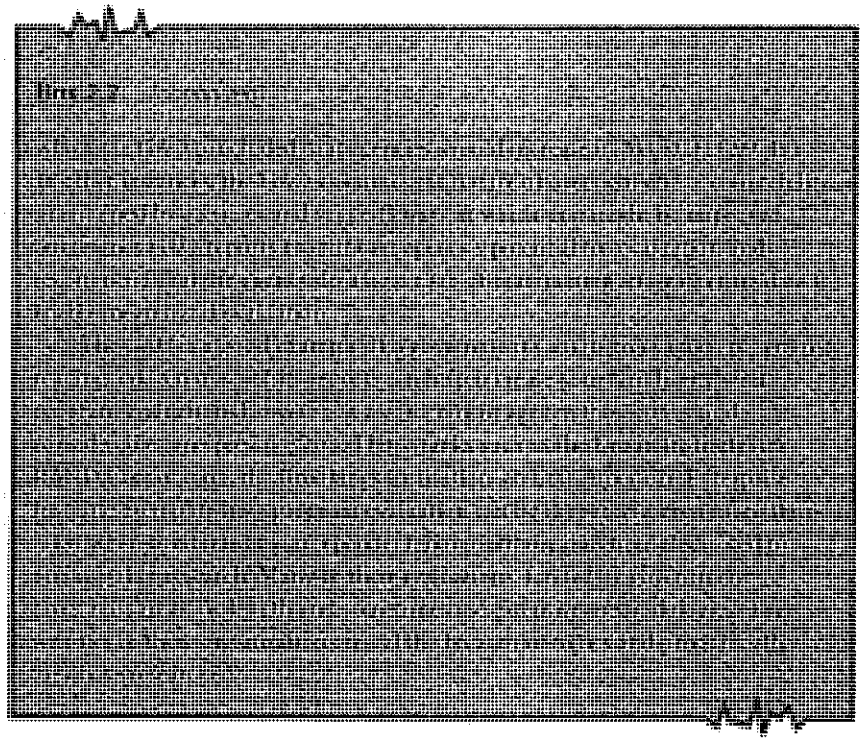
Box 2.2 Program Evaluation in Practice

Program evaluations are undertaken by many organizations, some of which emphasize and excel in this activity. The Government Accountability Office (GAO, formerly the General Accounting Office) of the U.S. Congress provides numerous assessments each year and is highly respected. Some GAO reports evaluate program outcomes. For example, a GAO study of 23 adult drug court programs found that recidivism rates were 10–30 percent lower among participants than among those in comparison groups (GAO-05-219). In this study, a comparison group was developed for each of the adult drug court programs. Some comparison groups were contemporaneous, that is, consisting of defendants who were eligible for the adult drug court program but who received conventional case processing. Other, historical comparison groups were developed from individuals who completed conventional case processing before the adult drug court was implemented. In each case, comparison group participants were selected to closely match characteristics of those in the adult court groups regarding substance abuse, socioeconomic status, demographic profile, and criminal justice history. However, recognizing the possibility of selection bias (matching is not perfect), the study used statistical methods to control for individual differences between adult court and comparison group members.

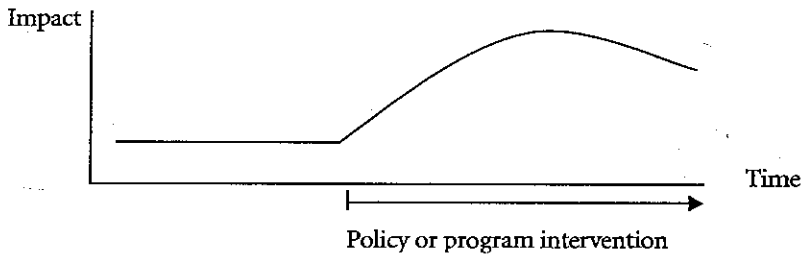
In another study, the GAO examined how long it took the Departments of State and Commerce to issue export licenses, which are required for exporting equipment and services that have military applications. The State Department issues licenses for items that have only military applications, and the Commerce Department issues licenses for items that have both military and commercial uses. Each year these agencies receive, respectively, 46,000 and 11,000 license applications (GAO-01-528). The study found little difference between these agencies; the State Department took 46 days to review an application, and the Commerce Department took 50 days. In making the comparison, the GAO was mindful to consider the nature and complexity of the application as a source of possible variation. Examples of GAO studies can be found at www.gpoaccess.gov/gaoreports/index.html.

However, many GAO reports are only descriptive, for example, providing information on what programs are doing, and focusing on issues of critical importance to the programs. For instance, a study of homelessness programs described what these programs do and examined the extent of coordination,

(continued)



The efficacy of the intervention might be suggested by a change in trend after the sixth month, controlled for any intervening confounding variables, of course. One such possibility is suggested below:



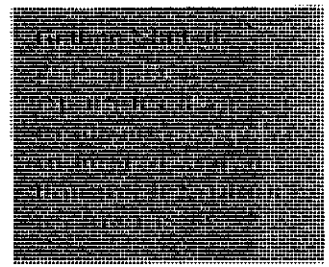
The additional observations, before and after the start of the intervention, allow analysts to research important questions on the persistence of intervention impacts. In our example, the impact of anger management programs might reach a saturation point, beyond which further decreases in high school violence are not observed. This would indicate that some causes of high school violence are not related to anger. Or students might

adopt new behaviors designed to mitigate or overcome the impact of anger management principles and practices. In this case, violence might begin to increase again after an initial decrease. Finally, how does the level of outcomes develop after the program ceases? Do students incorporate the new behaviors permanently, or do they go back to their old ways? A variety of such policy impact models can be examined toward the end of the intervention, when time series data are available.

The before-and-after design with a comparison group (design B1) combines features of designs B3 and B4 and resembles the classic, randomized experiment, except that subjects are not randomly assigned to the experimental and comparison groups. For example, in a study of innovative housing vouchers or mental health interventions, subjects with similar conditions might be assigned to different programs in similar cities that have similar neighborhoods. Some subjects might be followed who do not participate in any program; these subjects would make up the comparison group. Others might be assigned to a traditional program and others to one or more innovative programs. Outcomes would then be compared across these groups, controlling for rival hypotheses and differences that might exist among the different populations and local conditions. Such a design is obviously quite extensive and often expensive to undertake; it requires the development and coordination of programs. Few evaluations are of this kind, but they provide excellent insight into the rigors of evaluation methodology.

The models described here can help in developing designs for specific program evaluations; they draw attention to the roles of comparison groups and of baselines that might be used.

Finally, considerable thought exists about *types* of rival hypotheses. Rival hypotheses may still arise either from substantive matters of the program or from the use of quasi-experimental designs. While not all of the following concerns or categories are likely to be important in every situation, they assist analysts in identifying those concerns that are most salient as rival hypotheses in their specific situations. The following discussion draws attention to issues that might otherwise be overlooked.



Threats to external validity are defined as those that jeopardize the generalizability of study conclusions about program outcomes to other situations. For example, suppose we evaluate one anger management program in one school setting and, on the basis of that evaluation, wish to generalize our study conclusions to anger management programs in all schools? Such a generalization might be invalid if conditions in these other schools differ from conditions in the school that was studied. Or perhaps there is

something unique about the study population or about the students in the other schools that makes generalization problematic. If generalization is a study objective, then such concerns should be considered during the study design phase; we would need to choose program settings that can be generalized to other settings.

Threats to internal validity are those that jeopardize the study conclusions about whether an intervention in fact caused a difference in the study population. These threats often question the logic of study conclusions. Many different types of such threats exist, which are sometimes referred to using the following categories. *History* refers to events that are not part of the intervention yet occur during the intervention and affect study outcomes. For example, a shooting rampage among high school students elsewhere might temporarily reduce violence in other schools as that event is discussed and digested. Hence, history might explain the study outcomes. *Maturation* refers to the natural development of subjects in ways that affect study outcomes but that are not affected by the intervention. For example, students may themselves learn to control their anger, apart from any anger management program. People do grow up. Did the program control for this possibility? *Testing* refers to subjects changing their behavior because they are being tested rather than because of the intervention. For example, if high school violence is measured partly by asking students how many episodes of anger they experienced recently (however defined), asking them this question may cause some students to view anger as a problem, and they might take steps to reduce it. Any reduction in violence is then caused by the act of being tested rather than by the intervention itself.

Instrumentation refers to changes in outcomes resulting from the way in which an instrument (such as a survey) measures the outcomes. Perhaps violence is measured partly by observation, and observers become more attuned to different forms of violence over time. This will inflate later (for example, post-intervention) measures of violence; the instrument (observation) measures more violence over time, regardless of whether more violence is occurring. *Statistical regression* refers to the fact that extreme scores tend to become less extreme over time; they regress toward the average. If we start out with students who all exhibit extreme violence of the worst kind, then it may not be possible for them to become any worse. Regardless of the intervention, the group is likely to improve. *Selection bias* refers to the problem that subjects may not be truly comparable between the experimental (intervention) and comparison groups. For example, the experimental group might have more at-risk students, as discussed earlier. *Mortality* refers to biases due to attrition of study subjects, for example, the transfer of angry students to other schools during the intervention. This event will reduce the incidence of violence, of course. *Imitation* occurs when some subjects in the comparison group learn of the intervention in

the experimental group and begin imitating such behavior. In our example, students in classes that do not receive the anger management intervention might also see reductions in violence as a result of students talking with each other. *Rivalry* occurs when subjects in the experimental and comparison groups begin competing with each other. In our example, these students might compete with each other for being the least (or most!) violent.[6,7]

The point of the above discussion is to identify rival hypotheses that take away from the credibility of study conclusions later. These concerns need to be identified before program evaluation, so that program evaluation can be designed to address them. Most program evaluations involve at least a few of the above threats, as well as other rival hypotheses that relate to substantive matters of the research topic. Then analysts address these concerns either through program design or by gathering data about them so that they, as control variables, can later be accounted for through statistical techniques. Indeed, data often can be gathered about the above phenomena, such as behaviors that study subjects engage in during the program or the impact of intervening events. In short, analysts should identify important rival hypotheses and find ways to address them. By identifying and addressing rival hypotheses at an early stage, analysts can hope to increase the validity and acceptance of their work.[8]

SUMMARY

Social science research methods are often applied to many problems of management and analysis. Analysis typically involves a range of qualitative and quantitative research methods, and both basic and applied research focuses. Program evaluation is an example of such an application.

Variables, defined as observable phenomena that vary, represent a cornerstone concept in scientific research. Programs and policies usually attempt to affect variables in some ways (for example, by increasing or decreasing some social or economic conditions), and analysis often involves studying these changes.

Relationships in social science are distinguished by whether they are probabilistic (occurring sometimes) or deterministic (occurring each time). Relationships in social science are often probabilistic. Relationships are further distinguished as being either causal or associational. Causal relationships show cause and effect. When relationships are causal, independent and dependent variables can be distinguished. Independent variables are variables that cause an effect on other variables, and dependent variables are variables that are affected by other variables. Programs and policies are commonly conceptualized as independent variables causing changes in dependent variables, or outcomes.

The purpose of program evaluation often is to establish the effect of programs or policies on outcomes. A common concern is that other factors in addition to the program or policy (for example, events or processes) also affect outcomes. These alternative explanations for outcomes are referred to as rival hypotheses, and the variables associated with them are called control variables. The analytical task is to identify these rival hypotheses, collect data for the control variables, and use statistical methods to take their impact into account.

Program evaluation often uses quasi-experimental research designs. Such designs typically use comparison groups and baseline measurement in a variety of ways. The theory of quasi-experimental research design includes consideration of different types of rival hypotheses, which are distinguished as threats to internal or external validity. Familiarity with these categories can help analysts to identify rival hypotheses that are salient to their specific program evaluation.

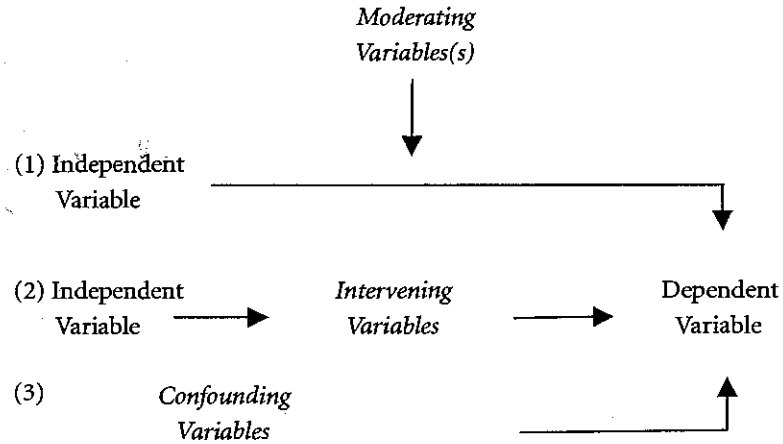
KEY TERMS

(includes bolded terms in the Section II introduction)

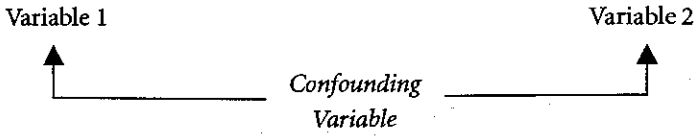
Applied research (see section introduction) (p. 17)
Association (p. 25)
Attributes (p. 21)
Basic research (see section introduction) (p. 17)
Causal relationships (p. 23)
Classic, randomized experiments (p. 29)
Constant (p. 21)
Control variables (p. 29)
Criteria for causality (p. 24)
Dependent variables (p. 23)
Descriptive analysis (p. 22)
Experimental designs (p. 29)
Hypotheses (p. 25)
Independent variables (p. 23)
Program evaluation (p. 25)
Qualitative research methods (see section introduction) (p. 17)
Quantitative research methods (see section introduction) (p. 17)
Quasi-experimental designs (p. 30)
Relationships (p. 22)
Research methodology (see section introduction) (p. 16)
Rival hypotheses (p. 29)
Six steps of program evaluation (p. 27)
Strategy of statistical control (p. 30)
Threats to external validity (p. 35)
Threats to internal validity (p. 36)
Variables (p. 21)

Notes

1. Later, in Chapter 9, we provide a more specific, technical definition of level of confidence, in our discussion of statistical significance.
2. Consider another example. In Louisiana a relationship exists between the increase in population and the state's shrinking in size each year. Does this mean that the weight of more people is causing the state to sink? No. The shrinking size is caused by erosion of coastal wetlands, not the weight of more people. These are unrelated events; the relationship is an association, only.
3. Two concerns are sometimes raised: (1) that the independent variable must precede the dependent variable in time and (2) that neither is caused by other variables; see the discussion of *spurious relationships* in note 5.
4. A variety of books address program evaluation. The leading text is Peter Rossi et al., *Evaluation: A Systematic Approach,* 7th ed. or later (Thousand Oaks, Calif.: Sage, 2003). See also *Exercising Essential Statistics,* the workbook that accompanies this textbook, for other references and exercises.
5. Control variables can affect the relationship between the independent and dependent variables in several ways. Different authors use different names to indicate these effects, but shown here is one approach. All three examples involve control variables. In the following graphic, moderating variables affect the way in which the independent variable affects the dependent variable, for example, sabotaging a class in which the instructor helps students learn to control their anger.



Sometimes a variable gives rise to two variables, when in fact no relationship exists between the two variables. This is called a spurious relationship, as is shown below. For example, the time of year (spring) is a spurious variable that gives rise to both storks and childbirths.



Bottom line: Regardless of how control variables affect variables, the point is to identify rival hypotheses (control variables) that may affect study conclusions.

6. A useful acronym for remembering these threats to internal validity is "Mis Smith:" maturation, instrumentation, selection, statistical regression, mortality, imitation, testing, and history. The classic source for these distinctions is Donald Campbell and Julian Stanley, *Experimental and Quasi-experimental Designs for Research* (Chicago: Rand McNally), 1963.
7. Even the classic, randomized research design is subject to some of these validity threats. Threats to external validity (generalizability) are a problem in any setting, and problems of history and mortality also may be present. It is not a given that the experimental and control groups experience the same intervening events (history), and they may have different rates of attrition (mortality). Testing might affect both groups, too. To address the problem of testing, a modification of the classic, randomized research design is the Solomon four-group design:

	Pretest	Program	Posttest
Group 1:	R O1	X	O2
Group 2:	R O3		O4
Group 3:	R	X	O5
Group 4:	R		O6

In this design, groups 3 and 4 allow the researcher to control for the impact of pretesting on groups 1 and 2.

8. For example, in a simple posttest design with no comparison group, the evaluation of anger management should consider whether any intervening effects (history) occurred that could have affected students' levels of anger and violence. Analysts will want to examine and account for possible maturation, statistical regression, and sample bias effects on the results. They also will want to assess subjects' knowledge of other efforts used elsewhere (which could give rise to imitation or rivalry) and ensure that the assessment method is accurate (minimizing effects of instrumentation).

Conceptualization and Measurement

CHAPTER OBJECTIVES

After reading this chapter, you should be able to

- Appreciate the challenge of measuring abstract concepts
- Implement methods for measuring abstract concepts
- Distinguish between different levels of measurement
- Apply a variety of Likert scales
- Create index variables
- Understand criteria for assessing measurement validity

Measurement is a foundation of science and knowledge. How well phenomena are measured affects what we know about them, and rigor in measurement increases the validity of analytical work. This chapter discusses key concepts of measurement and shows how to apply these measurements in analytical work such as program evaluation. This chapter also shows how to make index variables.

MEASUREMENT LEVELS AND SCALES

A ***scale*** is defined as the collection of attributes used to measure a specific variable. For example, the variable "gender" is commonly measured on a