CHAPTER 3

RESEARCH DESIGNS FOR PROGRAM EVALUATIONS

Introduction 81

What Is Research Design? 83
   The Origins of Experimental Design 84

Why Pay Attention to Experimental Designs? 89

Using Experimental Designs to Evaluate Programs:
The Elmira Nurse Home Visitation Program 91
   The Elmira Nurse Home Visitation Program 92
   Random Assignment Procedures 92
   The Findings 93
   Policy Implications of the Home Visitation Research Program 94

Establishing Validity in Research Designs 95

Defining and Working With the Four Kinds of Validity 96
   Statistical Conclusions Validity 97
   Working with Internal Validity 98
      Threats to Internal Validity 98
   Introducing Quasi-Experimental Designs:
      The Connecticut Crackdown on Speeding and the
      Neighborhood Watch Evaluation in York, Pennsylvania 100
Over the last 30 years, the field of evaluation has become increasingly diverse in terms of what is viewed as proper methods and practice. During the 1960s and into the 1970s most evaluators would have agreed that a good program evaluation should emulate social science research and, more specifically, methods should come as close to randomized experiments as possible. The ideal evaluation would have been one where people would have been randomly assigned either to a program or to a control group, the program was implemented, and after some predetermined period of exposure to the program, quantitative comparisons were made between the two groups. Program success was tied to statistically significant differences between program and control group averages on the variable(s) of interest.

Large-scale social experiments were funded, and evaluations were set up that were intended to determine whether the programs in question produced the outcomes that their designers predicted for them. Two examples of such experiments were the New Jersey Negative Income Tax Experiment and the Kansas City Preventive Patrol Experiment. In the New Jersey experiment, samples of low income families were randomly assigned to treatment groups where each group received a combination of minimum guaranteed family income plus a specific negative income tax rate. The latter worked like this: A family that was above the poverty line (say, $10,000 of income a year) would receive no benefit from the negative income tax part of the experiment. But if the family income fell below $10,000 per year, there would be payments that were related to how far below $10,000 the family earned. The lower the family income, the greater the payment—the greater the negative income tax (Pechman & Timpane, 1975).

The Kansas City Preventive Patrol Experiment (Kelling, 1974) was intended to test the hypothesis that the level of routine preventive patrol in a neighborhood would not affect the actual crime rate (measured by surveys of residents), the reported crime rate, or citizen perceptions of safety and security (measured by surveys of residents). Police beats were randomly assigned to one of three conditions: no routine preventive patrol (police would only enter the beat if there was a call for their services); normal levels of patrol; and two to three times the normal level of patrol. The experiment was run for a year, and during that time extensive measurements of key variables were made. The designers of the experiment believed that if the level of patrol could be shown to not affect key crime and citizen safety indicators, police departments across the United States could save money by modifying the levels of patrol that they deployed.
Typically, the programs were administered to samples of clients chosen randomly from a specific population, and clients for the control groups were chosen in the same way. The dominant belief in this approach was rooted in the assumption that randomized experiments are the best way to determine what effects programs have—they are the best test of the cause and effect relationship between the program and its intended outcomes.

In the 1970s and into the 1980s, as the field of evaluation diversified, some evaluators introduced and elaborated the use of quasi-experiments as ways of assessing cause and effect relationships in evaluations (Cook & Campbell, 1979). The costs and administrative challenges of doing randomized experiments prompted an interest in research designs that retained some of the features of experiments, but generally were more feasible to implement.

But, as we will see in Chapter 5, the biggest single change in the field during the 1970s and beyond was the introduction of qualitative program evaluations. The key difference between advocates of qualitative evaluations (Guba & Lincoln, 1981; Patton, 1997) and those who continued to assert the superiority of experiments and quasi-experiments was the qualitative evaluators’ emphasis on words as a basis for the analyses in evaluations. Advocates of experiments and quasi-experiments tended to emphasize the use of quantitative techniques, often involving applications of statistics to numerical data.

Today, we continue to have this diversity of approaches to evaluation. As you can see just by reviewing the chapter titles and topics in this book, program evaluators can choose from a wide range of tools, representing differing views of the field. Knowing which ones are appropriate in a given situation is one of the skills that practitioners must master.

Although there are many different reasons for conducting evaluations, a principal one is to learn whether the program achieved its intended outcomes. As we saw in Chapter 1, the question of program effectiveness really is two separate questions:

Was the program responsible for (or the cause of) the observed outcomes?

Were the observed outcomes consistent with the expected outcomes?

In Chapter 2, we learned how to describe programs as logic models, and in doing so, how to make explicit, the intended causal linkages in the program process. In this chapter, we focus on ways of testing the intended causal linkages in the program, beginning with the principal one: the connection between the program and the observed outcomes.
WHAT IS RESEARCH DESIGN?

Research design is fundamentally about examining the linkage depicted in Figure 3.1.

Notice what we have done in Figure 3.1. We have taken the program, which we “unpacked” in Chapter 2 in logic models, and repacked it. The complexity of logic models has been simplified again so that the program is back in a box.

Why have we done this? Wouldn’t it make more sense to keep the logic models we have worked on so far and test the causal linkages in such models? That way, we would be able to confirm whether the intended linkages between various outputs, linking constructs, and outcomes are supported by evidence gathered in the evaluation.

The reason we “repack” the logic models is that research designs are actually fairly blunt instruments in the tool kit of an evaluator. Research designs demand that we isolate each causal linkage that we wish to test. In order to examine a given cause and effect relationship, we must find ways of testing it while holding constant other factors that could influence it. A typical program logic will have a number of important causal linkages. In order to test these linkages using research designs we would need to isolate each one in turn to know whether that particular linkage is supported by evidence, holding constant the linkages in the rest of the logic model.

The problem is in finding ways of holding everything else constant while we examine each linkage in turn. In nearly all evaluations of programs, we simply do not have the time or the resources to do it. Thus, in thinking about research designs, we tend to focus on the main linkage, that being the one between the program as a whole (back in its box) and the observed outcomes.

Later in this chapter, we look at ways that have been developed to more fully test program logics. One approach that is quite demanding in terms of resources is to conduct an evaluation that literally tests all possible combinations of program components in an experimental design (Cook & Sciolli, 1972). Another one that is more practical is to use several complementary
research designs in an evaluation, and test different parts of the program logic with each one. These designs are often referred to as patched-up research designs (Poister, 1978), and usually, they do not test all the causal linkages in a logic model.

Research designs that fully test the causal links in logic models often demand more resources than are available to an evaluator. We will look at an example of such a research design later in this chapter to see what is involved in fully testing a logic model.

The Origins of Experimental Design

Experimental design originated in disciplines where it was essential to be able to isolate hypothesized cause and effect relationships. In agricultural research in the post–World War I period, for example, people were experimenting with different kinds of seeds to produce higher yields. There was keen interest in improving crop yields—this was a period where agriculture was expanding and being mechanized in the United States and Canada.

Researchers needed to set their tests up so that variation in seed types was the only factor that could explain the number of bushels harvested per acre. Typically, plots of a uniform size would be set up at an agricultural research station. Care would be taken to ensure that the soil type was uniform across all the plots and was generalizable to the farmlands where the grains would actually be grown.

Then, seed would be planted in each plot: carefully controlling the amount of seed, its depth, and the kind of process that was used to cover it. Again, the goal was to ensure that seeding was uniform across the plots. Fertilizer may have been added to all plots (equally) or to some plots to see if fertilizers interacted with the type of seed to produce higher (or lower) yields.

The seed plots might have been placed side by side or might have had areas of unplanted land between each. Again, that might have been a factor that was being examined for its effects on yield.

During the growing season, moisture levels in each plot would be monitored, but typically, no water would be provided other than rainfall. It was important to know if the seed would mature into ripe plants with the existing rainfall in that region. Because the seed plots were in the same geographic area, it was generally safe to assume that rainfall would be equal across all the plots.

Depending on whether the level of fertilizer and/or the presence of unplanted land next to the seed plots were also being deliberately manipulated along with the seed type, the research design might have been as
simple as two types of plots: one type for a new “experimental” seed, and the other for an existing, widely used seed. Or, the research design might have involved plots that either received fertilizer or did not, and plots located next to unplanted land or not.

Figure 3.2 displays a research design for the situation where just the seed type is being manipulated. As a rule, the number of plots of each type would be equal. As well, there would need to be enough plots so that the researchers could calculate the differences in observed yields and statistically conclude whether the new seed improved yields. Statistical methods were developed to analyze the results of agricultural research experiments. Ronald A. Fisher, a pioneer in the development of statistical tools for small samples, worked at the Rothansted Experimental [Agricultural] Station in England from 1919 to 1933. His book, *Statistical Methods for Research Workers* (Fisher, 1925) is one of the most important statistics textbooks written in the 20th century.

In Figure 3.2, “X” denotes the factor that is being deliberately manipulated, in this case, the seed type. More generally, the “X” is the treatment or program that is being introduced as an innovation to be evaluated. The $O_1$ and $O_2$ are observations made on the variable that is expected to be affected by the “X.” Treatments or programs have intended outcomes. An outcome that is translated into something that can be measured is a variable. In our case, $O_1$ and $O_2$ are measures of the yield of grain from each group of seed plots: so many bushels per acre (or an average for each group of plots).

Figure 3.3 displays the more complex research design involved when seed type, fertilizer, and cultivated (nonseeded) land nearby are all being manipulated. Clearly, many more seed plots would be involved—costing considerably more money to seed, monitor, and harvest. Correspondingly, the amount of information about yields under differing conditions would be increased.

Figure 3.3 is laid out to illustrate how the three factors (seed type, fertilizer, contiguous cultivated land) that are being manipulated would be “paired up” to fully test all possible combinations. In each of the cells of the
figure, there are the original two types of plots: those with the new seed and those without. The plots where “X” has occurred in each of the four cells of the figure have the new type of seed, and the plots in each of the four cells that do not get “X” are planted with regular seed. Because each cell in Figure 3.3 represents a different treatment, each “X” has been subscripted uniquely. In effect, the simpler research design illustrated in Figure 3.2 has been reproduced four times: once for each of the combinations of the other two factors.

This agricultural experimental research design and others like it have been generalized to a wide range of program evaluation situations. In program evaluations, research designs work best where the evaluator is involved in both the design and implementation of a program, and there are sufficient resources to achieve a situation where the effects of the program can be determined, while other factors are held constant. The paradigmatic situation is an experiment where, for example, clients are randomly assigned to the groups that do and do not get the program, and differences in client outcomes are measured. If the experiment has “worked,” outcome differences can confidently be attributed to the program: we can say that the program caused the observed difference in outcomes, that is, the causal variable occurred before the observed effect, the causal variable co-varied with the effect variable, and there were no plausible rival hypotheses.

The difficulty in most program evaluation settings, of course, comes in ruling out the possible rival hypotheses. Lawrence Mohr (1995) makes this point when he states: “The crux of the analysis of the efficacy of a treatment or program with respect to a particular outcome . . . is a comparison of what did
appear after implementing the program with what would have appeared had the program not been implemented. Events in the what-would-have-happened category must obviously be troublesome” (p. 4, italics added).

Figure 3.4 suggests a visual metaphor for what research design strives to achieve. The causal linkage between the program and the observed outcomes is isolated so that other, plausible rival hypotheses are deflected. The line surrounding the program and its outcomes in the figure represents the barrier against rival hypotheses which is created by a defensible research design.

Achieving this isolation of the program/observed outcomes linkage is often associated with designing and implementing the program as a randomized experiment. Table 3.1 shows two different experimental designs: the first involves measuring the outcome variable(s) before the program is implemented, and then after. This design is **before-after design** is the classic experimental design and is often used where evaluators have sufficient resources and control to design and implement a before-after outcome measurement and data collection process. Having done so, it is possible to calculate the before-after changes in both the treatment and control groups, as well as ensuring that the pretest results indicate the two groups are similar before the program begins. The second design is called the **after-only experimental design** and does not measure program/control differences before the treatment begins; it generally works well, where random assignment to treatment and control groups has occurred. Random assignment generally assures that the only difference between the two groups is the treatment. Both designs in Table 3.1 include “no program” groups to achieve the “all other things being equal” comparison, which permits us to see what differences, if any, the program makes.

The random assignment of cases/units of analysis to treatment and control groups is indicated in Table 3.1 by the letter “R” in front of both the treatment and control groups. This process is intended to create a situation where the clients in the program and no-program groups are equivalent in
all respects, except that one group gets the program. Not pretesting the two groups assumes that they are equivalent. Where the numbers of clients randomly assigned are small, pretesting can establish that the two groups are really equivalent, at least in terms of the outcome measure(s).

When reviewing Table 3.1, keep in mind that the “X” designates the program or treatment of interest in the evaluation. The “O”s indicate measurements of the outcome variable of interest in the evaluation. The “R” indicates that we have assigned cases/units of analysis (usually people) randomly to the two groups. Where we measure the outcome variable before and after the program has been implemented, we are able to calculate the average change in the level of outcome. For example, if our program was focused on improving knowledge of parenting skills, we could measure the average gain in knowledge (the “after” minus the “before” scores) for the program group \((O_2-O_1)\) and the control group \((O_4-O_3)\). When we compare the average gain in outcome levels between the two groups after the program has been implemented, we can see what the incremental effect of the program was.

Where we do not have pretest measures of the outcome variable, as in the after-only experimental design, we would compare the averages of the program and control groups after the program was implemented \((O_1-O_2)\). One additional thing to keep in mind is that in experimental evaluations where we have several outcomes of interest, we have separate (but parallel) experimental designs for each variable. For example, if we were interested in evaluating the attitude changes towards parenting as well as the knowledge gains from a parenting program, we would have two research designs—one for each outcome. Likewise, where we have more than one treatment, or combination of treatments, each combination is designated by a separate X, subscripted appropriately.

---

**Table 3.1 Two Experimental Designs**

<table>
<thead>
<tr>
<th>Pre-Test Post-Test Design (Classic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R₁  O₁   X   O₂</td>
</tr>
<tr>
<td>R₂  O₁   O₄</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Post-Test Only Design (After-Only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R₁   X    O₃</td>
</tr>
<tr>
<td>R₂    O₂</td>
</tr>
</tbody>
</table>
The second design in Table 3.1 does have the potential of solving one problem that can arise in the first design. Sometimes, pretesting can be intrusive and have its own effect on the posttest measurement. Suppose you are evaluating a server intervention program that is intended to train employees who serve alcoholic beverages in bars and other establishments. One expected outcome might be improved knowledge of ways to spot customers who should not be served any more drinks, and how to say “no” in such situations.

If “knowledge level” was measured with the responses to a set of true-false statements before the training, and the same instrument was used after the training, we might expect higher scores simply because employees are familiar with the instrument. Using the first design might produce outcomes that are higher than they should be, misleading those who might want to generalize the results to other program situations.

WHY PAY ATTENTION TO EXPERIMENTAL DESIGNS?

The field of program evaluation continues to debate the value of experimental designs (Shadish, Cook, & Campbell, 2002). On the one hand, they are generally seen to be costly, require more control over the program setting than is feasible, and are vulnerable to a variety of implementation problems. But for some evaluators, and some government jurisdictions, experimental designs continue to be “the gold standard” when it comes to testing causal relationships (Lipsey, 2000; Office of Management and Budget, 2004).

Weisburd (2003), in a discussion of the ethics of randomized trials, asserts that the superior (internal) validity of randomized experiments makes them the ethical choice in criminal justice evaluations.

At the core of my argument is the idea of ethical practice. In some sense, I turn traditional discussion of the ethics of experimentation on its head. Traditionally, it has been assumed that the burden has been on the experimenter to explain why it is ethical to use experimental methods. My suggestion is that we must begin rather with a case for why experiments should not be used. The burden here is on the researcher to explain why a less valid method should be the basis for coming to conclusions about treatment or practice. The ethical problem is that when choosing non-experimental methods we may be violating our basic professional obligation to provide the most valid answers we can to the questions that we are asked to answer. (p. 350)
Although Weisburd’s view would be supported by some advocates of experimentation in evaluation, most practitioners recognize that there can be risks associated with randomized experiments. Shadish, Cook, and Campbell (2002) discuss the ethics of experimentation and point out that in the history of research with human participants, there are examples that have shaped our current emphasis on protecting the rights of individuals, including their right to informed consent, before random assignment occurs.

Among the best known American examples of controversial experiments was a series of studies conducted by Stanley Milgram (1983). His work focused on obedience to authority and was intended to see whether American university students who had agreed to be experimental subjects could be induced to behave in ways that would injure others, based on obeying the orders of the experimenter. Milgram demonstrated that his subjects regularly behaved in ways that they believed was harming a person whom they were order to punish in the experiments.

One feature of the obedience experiments was deception: the subjects did not know that the experimenter was colluding with the person who was supposedly being punished. Instead the subjects thought that they were actually administering electric shocks that were harmful or perhaps even fatal.

Deception has become a central concern with any research, but is highlighted in situations where people are randomly assigned and one group does not receive a possible benefit. In situations where the participants are disadvantaged (socially, economically, or mentally), even informed consent may not be adequate to ensure that persons fully understand the consequences of agreeing to random assignment. Shadish, Cook, and Campbell (2002) suggest strategies for dealing with situations where withholding treatment is problematic. For example, persons assigned to the control group can be promised the treatment at a later point.

We discuss the ethics of evaluation in Chapter 11 and suggest core values that need to be taken into account in conducting evaluations. The evaluation profession has guidelines for practice, but unlike most other professions, they are not enforceable.

Some program evaluators have argued that because opportunities to use experimental or even quasi-experimental designs are quite limited, the whole idea of making experiments the paradigm for program evaluations that examine causes and effects is misleading. We are setting up an ideal that is not achievable and expecting evaluators to deal with issues that they cannot be expected to resolve. As Berk and Rossi (1999) argue, “there is really no such thing as a truly perfect evaluation, and idealized textbook treatments of research design and analysis typically establish useful aspirations but unrealistic expectations” (p. 9).
The reality is that many situations in which evaluations are wanted simply do not permit the kind of control and resources that experiments demand, yet we do proceed with the evaluation, knowing that our findings, conclusions, and recommendations will be based in part on evidence that does not meet the standards implied by the experimental approach. Evidence is the essential core around which any program evaluation is built, but the constraints on resources and time available will usually mean that at least some issues that ideally should be settled with data from experiments will, in fact, be settled with professional judgments.

The key point is that there is value in knowing the requirements for experimental evaluations. Understanding the logic of experimental designs is important for understanding what we are claiming when we address issues of program effectiveness. Not being able to conduct an experiment does not change the need for us to appreciate what is involved in answering the question: Did the program cause the outcomes that we observe? We believe that most evaluators who are asked to address the issue of program effectiveness can tackle this question—perhaps not answering it definitively, but instead, reducing the uncertainty around the question.

**USING EXPERIMENTAL DESIGNS TO EVALUATE PROGRAMS:**

**THE ELMIRA NURSE HOME VISITATION PROGRAM**

In the field of evaluation, there is a rich literature that chronicles the experiences of researchers and practitioners with studies where a core feature is the use of randomized experiments. Although the field has and continues to diversify in terms of criteria for judging appropriate evaluation designs, randomized experiments continue to be a key part of our profession. In the United States, there are currently several groups of researchers who are committed to using randomized experiments to understand how programs work to ameliorate social problems.

One such group is headed by David Olds, and since 1977 he and his colleagues have conducted three extensive randomized experiments in Elmira, New York, Memphis, Tennessee, and Denver, Colorado to determine whether home visits by nurses to first-time mothers prevent poor pregnancy outcomes, prevent subsequent health and development problems, and improve the mother’s own life (Olds, Henderson, Kitzman, et al., 1998). The Elmira study was the first one conducted and we will summarize its main features to illustrate how randomized experiments have been used to address social policy issues.
The Elmira Nurse Home Visitation Program

Begun in 1977, the Elmira Nurse Home Visitation Program was implemented with pregnant women who had not given birth before and had one or more of the following characteristics: were young (19 years or younger), a single parent, or were poor. These conditions, or combinations of them, were expected to result in the mothers being more likely to be smokers, drug or alcohol users, come from families where their own childhood experiences were difficult, have diet and exercise deficiencies, and in general not have the health, knowledge, skills, and attitudes that predicted successful birthing and caring for their first baby.

Random Assignment Procedures

A total of 400 women were randomly assigned to one of four groups. In group 1, 94 women were offered screening services for their babies at age 1 and again at age 2 to see whether the babies had developmental problems. In group 2, 90 were offered the screening services plus free transportation to their prenatal and postnatal medical appointments, up to age 2. In group 3, 100 women were offered screening, transportation and regular home visits by nurses during their pregnancy. And in group 4, 116 women got screening, transportation, and home visits both before they gave birth and for the first 2 years of their child’s life. The first two groups were the “control group” in the main analyses, and groups 3 and 4 were the “experimental group,” those mothers received home visits by nurses.

The randomization procedure occurred in several stages. First, women who generally fit the criteria for “at risk” mothers were invited to participate in the study. Between 1978 and 1980, 500 women were interviewed and of those, 400 agreed to participate in the study. Each participant was assigned to one of the four groups using a deck of cards—at the end of the intake interview, each woman was invited to draw a card from a deck that consisted of randomly shuffled cards with a number from 1 to 4 on each card. As the 400 participants were recruited, the deck was periodically rebalanced to overrepresent the groups that had fewer participants (Olds, Henderson, Chamberlin, & Tatelbaum, 1986).

Comparisons of the control and experimental (or program) groups at the point where they were recruited indicated that the two groups were equivalent in their background characteristics. Very few of the participants dropped out of the study before giving birth, meaning that the two groups continued to be equivalent in terms of background characteristics (Olds
et al., 1986). Table 3.2 shows the research design for the first 3 years of the Elmira study (prenatal to 2 years of age).

In the research design, there were many different variables to measure the effects of the treatments on the children and the mothers. These have been designated as the $O_{1-4}$ to $O_n$ series of variables for the four groups.

The Findings

The findings from the Elmira nurse home visitation experiment suggest that when the home-visited mothers are compared to those that did not get this program, home-visited mothers tended to change their attitudes and behaviors before giving birth: their diets improved; they smoked fewer cigarettes; they had fewer kidney infections; they took advantage of community services; and among young mothers aged 14 to 16, they tended to give birth to heavier babies.

In the first years of childhood, home-visited mothers tended to have fewer incidents of abused or neglected children, less injuries for their children, less

---

### Table 3.2 Research Design for the Elmira Nurse Home Visitation Program

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_1$</td>
<td>$X_1$</td>
<td>$O_1$ to $O_n$</td>
</tr>
<tr>
<td>$R_2$</td>
<td>$X_2$</td>
<td>$O_2$ to $O_n$</td>
</tr>
<tr>
<td>$R_3$</td>
<td>$X_3$</td>
<td>$O_3$ to $O_n$</td>
</tr>
<tr>
<td>$R_4$</td>
<td>$X_4$</td>
<td>$O_4$ to $O_n$</td>
</tr>
</tbody>
</table>

**Control Group**

- $X_1$ is developmental screening for the first born child, ages 0 to 2.
- $X_2$ is screening plus transportation for pre- and postnatal appointments.

**Program Group**

- $X_3$ is screening, transportation, and prenatal nurse home visits.
- $X_4$ is screening, transportation, prenatal nurse home visits, and postnatal home visits.
severe injuries where there was physical abuse, fewer subsequent pregnancies, and more work force participation.

Although the initial plan was to follow the mothers and their children up to age 2, additional funding permitted the study to continue so that the parents and children have been followed to age 15. The longer time frame has meant that additional variables could be measured: home visited mothers had fewer subsequent births; longer times until the next birth after their first born child; less time on welfare; and among the children, fewer behavior problems due to substance abuse and fewer arrests.

A cost-benefit analysis comparing the two groups up to age 15 showed that the costs of the program were considerably less than the benefits to society from problems (and costs) prevented by the program, particularly for families where the mother had been low income and single at the time she was pregnant. The experimental design made it possible to say that the problems prevented (arrests, for example) were attributable to the home visitation program.

Policy Implications of the Home Visitation Research Program

The Elmira study is part of a sustained research program that has been focused on the question of how to reduce child and family problems associated with young women in difficult circumstances getting pregnant and giving birth. Over the past 28 years, randomized controlled trials of different combinations of program components have been conducted by Olds and others in different locations in the United States (Olds, O'Brien, Racine, et al., 1998). The commitment of the evaluation community to this issue suggests that the cumulative findings should be the basis for public policies. However, even with the sustained research and cumulative evidence that home visits by nurses do work, there are substantial barriers to the widespread dissemination of these programs.

First, critics point out that the findings from studies are not consistent. Given the volume of research, the variability in methods and rigor, some studies find no effects of nurse home visits, notwithstanding that the most rigorous ones consistently find positive effects from this intervention (Olds, Hill, Robinson, et al., 2000). Public policy makers can exploit this diversity in findings by suggesting that the results are not strong enough.

Second, advocacy groups tend to oversell the research results to make their case that public monies ought to be spent. Again, the issue that is ultimately raised by this approach is the credibility of the programs.
Third, these programs are costly. And in a political environment where the trend is to limit or even reduce social programs, any efforts to spend more are met with skepticism.

Finally, even the most rigorous randomized controlled trials of these programs run into a substantial problem in being able to translate program components that are tailored to an experiment into institutionalized programs. The variability in local circumstances means that even if funding is available, fidelity to the program components that have been tested is not assured. Even in the Elmira program, once the 2-year point after the births of the first-born children had occurred, changes in the leadership of the local funding agency resulted in substantial changes to the focus of the program. It took considerable pressure and negative publicity to steer the program back to its original focus (Olds, O’Brien, Racine, et al., 1998).

Olds and his colleagues sum up their own stance about the policy relevance of their research in the following way (Olds, O’Brien, Racine, et al., 1998):

The divisiveness of current national policy debate and the understandable public skepticism about a social science which in the past has promised more than it can deliver, should send us neither into hiding while we wait for a better day nor into overzealous advocacy as we declare our political allegiances. Rather, we should seek to maintain both the integrity and utility of the scientific enterprise—both its capacity for producing new knowledge and the applicability of that knowledge to social problems that must be addressed in the present. This, at least, is the perspective we intend to adopt in our continuing effort to make what we learn through careful study about effective home visitation service relevant in the public domain. (p. 98)

ESTABLISHING VALIDITY IN RESEARCH DESIGNS

The example of spurious increases in knowledge due to testing and re-testing knowledge in a server-training program is one possible rival hypothesis that could serve as an alternate explanation for observed outcomes. Having a no-program group mitigates that problem, but it persists if efforts are made to use evaluation results as a benchmark for other, or future, program settings.

Over the last 35 years, several major contributions have been made to describe the ways that research designs can be weakened by validity problems. Cook and Campbell (1979) defined and described four different classes of validity problems that can compromise research designs in program evaluations.
Figure 3.5 shows the four kinds of validity described by Cook and Campbell, and suggests ways that they can be linked. **Statistical conclusions validity** "feeds into" **internal validity**, and the two together support **construct validity**. All three support **external validity**. The questions in Figure 3.5 indicate the key issue that each kind of validity is intended to address. Notice that statistical conclusions validity and internal validity focus on the variables as they are measured in a program evaluation. Construct validity and external validity are both about generalizing; the former involves generalizing from the measured variables back to the constructs in the program model, and the latter is about generalizing the evaluation results to other situations.

In the following section, we will expand on these four kinds of validity, and the threats to each kind.

**DEFINING AND WORKING WITH THE FOUR KINDS OF VALIDITY**

There are three conditions for establishing a causal relationship between two variables (Shadish et al., 2002).

**Temporal asymmetry**, that is, the variable that is said to be the cause, precedes the variable that is the effect.
Co-variation, that is, as one variable varies, the other also varies either positively or negatively.

No plausible rival hypotheses, that is, no other factors that could plausibly explain the co-variation between the two variables.

These three conditions are individually necessary and jointly sufficient for establishing a causal relationship between two variables. The first tends to be treated at a conceptual level as well as at an empirical level. We hypothesize temporal asymmetry, and then look for ways of observing it in our program implementations. The second and third conditions are addressed by statistical conclusions validity and internal validity, respectively.

Statistical Conclusions Validity

Statistical conclusions validity is about establishing the existence and strength of the co-variation between the cause and the effect variables. To do so, one must pay attention to issues that fall within the purview of statistical analysis.

This book is not concerned with the details of statistical methods. Indeed, there are a large number of statistical tests that program evaluators can use, each having its own requirements for defensible applications. The appendix to this chapter offers a succinct summary of some of the statistical methods that are used by evaluators, as well as conditions under which they can be used.

For the purposes of this book, it is important to know that statistical conclusions validity involves the following (Cook & Campbell, 1979):

- Ensuring that sampling procedures are adequate, if they are used to gather data
- Ensuring that the appropriate statistical tests have been applied and that the outcomes have been interpreted correctly
- Ensuring that the measures that are used are sufficiently reliable

The third condition focuses on a general assumption that is usually made in applying most statistical tests—the theoretical models that underlie statistical tests assume that the measurement procedures that were used to gather the data are reliable (Berry, 1993).
Working With Internal Validity

Internal validity is about ruling out rival hypotheses. Beginning with Campbell and Stanley (1966), contributors to the program evaluation literature have categorized threats to the internal validity of research designs. We will learn about nine kinds of threats to internal validity. Each one is a category of possible problems and in a given situation there may be none, one, or even several factors that are a problem for a particular category.

**Threats to Internal Validity**

1. **History**: External events that coincide with the implementation of a program.

   Example: A state-wide Counter Attack program, aimed at reducing accidents and injuries on state highways due to alcohol consumption is introduced in May. A state seat belt law is introduced in October of the same year. Because the seat belt law is intended to reduce accidents and injuries, it is virtually impossible to disentangle the outcomes of the two programs.

2. **Maturation**: As program participants grow older, their behaviors tend to change in ways that appear to be program outcomes.

   Example: A youth vandalism prevention program in a community is developed in a short period of time during a period of rapid population growth. The population matures roughly as a cohort. Children born in the community also mature as a cohort. If a program is developed to “combat a rising youth vandalism problem” when the average youth age is 12, by the time the average age is 16, the community may have outgrown the problem even without the program.

3. **Testing**: Taking the same posttest as had been administered as a pretest can produce higher posttest scores due to learning about the testing procedure.

   Example: Using a pre/postmeasure of knowledge in a server training program as described earlier, servers score higher on a test of “knowledge level” after the training, not because they have increased their knowledge during training, but simply because they are familiar with the test from when it was administered before the training.
4. **Instrumentation:** As the program is implemented, the way in which key variables are measured also changes.

Example: A program to decrease burglaries is implemented at the same time that the records system in a police department is automated: reporting forms change, definitions of different types of crimes are clarified, and a greater effort is made to “capture” all crimes reported in the database. The net effect is to “increase” the number of reported crimes.

5. **Statistical regression:** Extreme scores on a pretest tend to have more measurement error. Thus, if program participants are selected because they scored low or high on the pretest, their scores will tend to regress toward the mean of the scores for all possible participants on the posttest.

Example: People are selected for an employment skills training program on the basis of low scores on a self-esteem measure. On the posttest, their self-esteem scores increase.

6. **Selection:** Persons/units of analysis chosen for the program are different from those chosen for the control group.

Example: A program to lower recidivism among youth offenders selects candidates for the program from the population in a juvenile detention center. The candidates are selected in part because they are thought to be reasonable risks in a half-way house living environment. If this group was compared to the rest of the population in the detention center (as a control group), differences between the two groups of youths, which could themselves predict recidivism, might explain program outcomes/comparisons.

7. **Mortality:** People/units of analysis “drop out” over the course of the program.

Example: A program to rehabilitate chronic drug users may lose participants who would be least likely to succeed in the program. If the pretest “group” were simply compared to the posttest group, one could mistakenly conclude that the program had been successful.

8. **Ambiguous temporal sequence in the cause and effect variable:** It is not clear whether the key variable in the program causes the outcome, or vice versa.

Example: A program that is intended to improve worker productivity hypothesizes that by improving worker morale, productivity
will improve. The data show that both morale and worker productivity improve. But the program designers may well have missed that improved morale is not the cause, but the effect. Or there is a reciprocal relationship between the two variables such that improvements in morale will induce improvements in productivity which, in turn, will induce improved morale, which in turn, improves productivity.

9. **Selection-based interactions**: Selection interacts with other internal validity threats so that the two (or more) threats produce joint effects (additive and otherwise) on variables of interest.

Example: A program to improve reading abilities in a school district is implemented so that program classes are located in higher-income areas and control classes in lower-income areas. Tests are given (pre, post) to both groups and the findings are confounded not only by selection bias, but also by the fact that higher-income children tend to mature academically more quickly.

Each of the nine categories of internal validity threats suggests possible solutions to the particular problems, although designing a study to sort out reciprocal or ambiguous causation can be challenging (Shadish et al., 2002). To avoid the intrusion of history factors, for example, anticipate environmental events that could coincide with the implementation of a program and, ideally, deploy a control group so that the history factors affect both groups, making it possible to sort out the incremental effects of the program.

The difficulty with that advice, or corresponding “solutions” to the other eight types of problems, is in having sufficient resources and control over the program design and implementation to structure the situation to effectively permit a “problem-free” research design.

*Introducing Quasi-Experimental Designs:*
*The Connecticut Crackdown on Speeding and the Neighborhood Watch Evaluation in York, Pennsylvania*

Fundamentally, all research designs are about facilitating comparisons. Well-implemented experimental designs, because they involve random assignment of units of analysis to treatment and control groups, are constructed so that program and no-program situations can be compared “holding constant” other variables that might explain the observed differences in program outcomes. It is also possible to construct and apply research designs that allow us to compare program and no-program conditions in
circumstances where random assignment does not occur. These quasi-experimental research designs typically are able to address one or more categories of internal validity threats, but not all of them.

Resolving threats to internal validity in situations where there are insufficient resources or control to design and implement an experiment usually requires the judicious application of designs that reduce or even eliminate threats that are most problematic in a given situation. Usually, the circumstances surrounding a program will mean that some potential problems are not plausible threats. For example, evaluating a 1-week training course for the servers of alcoholic beverages is not likely to be confounded by matura-
tion of the participants.

In other situations, it is possible to construct research designs that take advantage of opportunities to use existing data sources, and combine these designs with ones involving the collecting of data specifically for the evaluation, to create patched-up research designs (Poister, 1978) that are combinations of quasi-experimental designs. Patched-up designs are usually stronger than any one of the quasi-experimental designs that comprise the patchwork, but typically still present internal validity challenges to evaluators.

The Connecticut Crackdown on Speeding

Donald T. Campbell and H. Laurence Ross (1968), in their discussion of the Connecticut Crackdown on Speeding, show how successive applications of quasi-experimental research designs can build a case for determining whether a program intervention is successful. They discuss three different ways of assessing whether the introduction of a state-wide crackdown on speeding motorists in 1955 resulted in a reduction of traffic-related fatalities.

The before-after design illustrated below shows that traffic fatalities in Connecticut had decreased from 325 in 1955 to 285 in 1956. In the diagram below, the “O” is reported traffic-related fatalities, measured before and after the crackdown on speeding was implemented. The crackdown was a policy change, not a program, and is the “X” in the diagram.

O X O

The reduction in fatalities could have been due to a number of other factors (threats to internal validity) including history (another policy or program might have been implemented then), maturation (automobile drivers on average might have been older and more safety-conscious), instrumentation (the way that reported accident statistics were gathered and recorded could have changed), and statistical regression (the reported accident level before the crackdown could have been unusually high, given the historical
trend). So, to reduce the likelihood of those threats, Campbell and Ross introduced a single time-series design, which compared fatality levels for 5 years before and 4 years after the crackdown in Connecticut. This time series can be diagramed as follows:

\[
\text{OOOOOXOOOO}
\]

The data for 9 years, displayed in Figure 3.6, suggested that maturation and instrumentation did not appear to be rival hypotheses. There is no overall trend downward that might have suggested that Connecticut drivers were maturing, hence, less likely to get take the risks associated with speeding. Although the before-after comparisons could have indicated that instrumentation was a problem, a review of data collection practices indicated that there was no change in the way accident data were gathered coincident with the crackdown. The data pattern displayed in Figure 3.6 could have been due to statistical regression—the jump in reported fatalities in 1954 was the biggest in the time series and would suggest the possibility of a movement in the next year back towards the mean of the pre-crackdown time series.
To deal with possible history threats (some other event or events that coincided with the crackdown) and to deal with statistical regression, Campbell and Ross looked at traffic fatalities in the four surrounding states (New Jersey, New York, Massachusetts, and Rhode Island), both individually and combined. This comparative time-series design suggested that the crackdown in Connecticut did reduce fatalities.

The comparison with nearby states can be diagrammed as follows:

![Diagram of traffic fatalities comparison]

**Figure 3.7** Reported Traffic Fatalities in Connecticut and Four Comparison States: 1951–1959 (per 100,000 population)


To deal with possible history threats (some other event or events that coincided with the crackdown) and to deal with statistical regression, Campbell and Ross looked at traffic fatalities in the four surrounding states (New Jersey, New York, Massachusetts, and Rhode Island), both individually and combined. This comparative time-series design suggested that the crackdown in Connecticut did reduce fatalities.

The comparison with nearby states can be diagrammed as follows:

```
OOOOOXOOOO
```

```
OOOOO OOOO
```

Figure 3.7 displays two time series: one for Connecticut and the one for the four control states (New Jersey, New York, Massachusetts, and Rhode Island), averaged together. The patterns in the data suggest that the drop in
reported fatalities in Connecticut was not replicated in the control states. So, although statistical regression continues to be a possible rival hypothesis, it needs to be weighted against the evidence from the control states. Further, the big jump in fatalities in Connecticut in 1954 also happened in Rhode Island (Campbell & Ross, 1968, p. 45), suggesting that large changes in year-over-year fatality levels were not unique to Connecticut.

Finally, Campbell and Ross examined the 9-year trends in several output variables in Connecticut—variables which might have been expected to be affected by the crackdown, and which would have contributed in turn to the intended outcome of reduced traffic fatalities. License suspensions for speeding increased sharply and speeding violations decreased. Both of these changes were consistent with expected program effects. In addition, an unintended effect was observed: arrests while driving with suspended licenses increased.

The York Neighborhood Watch Program

A second example of such a strategy was the evaluation of a neighborhood watch program in York, Pennsylvania, in the 1970s (Poister, McDavid, & Magoun, 1979). The program was intended to reduce crime, but was focused specifically on reducing burglaries. It was implemented in one area of the city and a no-program “control” area was established for comparison.

Reported burglaries were tracked at both the neighborhood and city-wide levels. In addition, a survey of all the block captains in neighborhood watch blocks was conducted to solicit their perceptions of the program, including estimates of resident attendance at neighborhood watch meetings. Finally, key environmental factors were also measured for the entire period, the principal one being the unemployment rate in the whole community.

Several research designs were embedded in the evaluation. At the level of the neighborhood watch blocks, the program was implemented and the block captains were interviewed.

\[ X \quad O \]

where \( X \) = the neighborhood watch program and \( O \) = the measurement of block captain perceptions of program activity. This research design is called an **implicit design** (sometimes called a **case study design**), which is a weak design by itself because it does not include any comparisons that control for possible threats to internal validity.

Reported burglaries were compared between the neighborhoods that received the program, and those that did not.

Program  OOOOOOOOXXXXOXXXXXOXXXXXOXXXXXO
No Program  OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO
where X = the neighborhood watch program and O = the reported burglaries in the program and no-program areas of the city. Notice that for the program area of the city, we show the “X”s and “O”s being intermingled. That shows that the program continued to operate for the full length of the time series, once it was implemented. This comparative time-series design is typically stronger than the implicit design because it includes a no-program group. Among the threats to this design is the possibility that the program group is not comparable to the no-program group (selection bias). That could mean that differences in reported burglaries are due to the differences in the two types of neighborhoods, and not the program.

Finally, reported burglaries were compared before and after the program was implemented, city-wide.

OOOOOOOOOOXOXOXOXOXOXOXO

This single-time series design is vulnerable to several internal validity threats. In this case, what if some external factor or factors intervened at the same time that the program was implemented (history effects)? What if the way in which reported burglaries were counted changed as the program was implemented (instrumentation)? What if the city-wide burglary rate had jumped just before the program was implemented (statistical regression)?

In this evaluation, several external factors (unemployment rates in the community) were also measured for the same time period and compared to the city-wide burglary levels. These were thought to be possible rival hypotheses (history effects) that could have explained the changes in burglary rates.

Findings and Conclusions From the Neighborhood Watch Evaluation

The evaluation conclusions indicated that at the block level, there was some activity, but attendance at meetings was sporadic. A total of 62 blocks had been organized by the time the evaluation was conducted. That number was a small fraction of the 300-plus blocks in the program area. At the neighborhood level, reported burglaries appeared to decrease in both the program and no-program areas of the city. Finally, city-wide burglaries decreased shortly after the program was implemented. But given the sporadic activity in the neighborhood watch blocks, it seemed likely that some other environmental factor or factors had caused the drop in burglaries.

Figure 3.8 displays a graph of the burglaries in the entire city from 1974 through 1980. During that time, the police department implemented two programs: a neighborhood watch program and a team-policing program. The latter involved dividing the city into team policing zones and permanently assigning both patrol and detective officers to those areas.
Figure 3.8 is divided into five time periods. The level of reported burglaries varies considerably, but by calculating a 3-month moving average (burglaries for January, February, and March would be averaged and that average reported for February; burglaries for February, March, and April would be averaged and that average reported for March, and so on), the graph is stabilized somewhat. The 3-month moving average is displayed as the dotted line.

As you can see by inspecting the graph, the police department initially implemented the neighborhood watch program, then shortly afterward moved to team policing as well. Both team-policing and the neighborhood watch program were in operation for period 3, then neighborhood watch was cancelled, but team policing continued (period 4). Finally, because the detective division succeeded in its efforts to persuade the department to cancel the team-policing program (detectives objected to being assigned to area-focused teams and wanted instead to operate city-wide), the police department restarted the neighborhood watch program (period 5).

Inspection of Figure 3.8 indicates that burglaries were increasing in the period prior to implementing the neighborhood watch program in 1976.
Burglaries dropped, but within 5 months of the neighborhood watch program being started up, team policing was implemented city-wide. When two programs are implemented so closely together in time, it is often not possible to sort out their respective outcomes—in effect, one program becomes a “history” rival hypothesis for the other. In this situation, the public response to a perceived burglary problem consisted of doing as much as possible to eliminate the problem. Although implementing the two programs may have been a good political response, it confounded any efforts to sort out the effects of the two programs, had the evaluation time frame ended in 1977.

By extending the time series as is shown in Figure 3.9, it was possible to capture two additional program changes: withdrawal of team policing in 1978 and the reinstatement of the neighborhood watch program at that point. Figure 3.9 depicts these program changes between 1974 and 1980. The neighborhood watch program is shown as a single-time series wherein the program is implemented (1976), withdrawn (1977–1978), and then implemented again (1978–1980). This on-off-on pattern facilitates being able to detect whether the program impacted on reported burglaries, notwithstanding some difficulties in drawing “borders” between the no-program and program periods. Because some neighborhood watch blocks could continue operating beyond the “end” of program funding in 1977, it is possible that some program outputs (block meetings, for example) for that program persisted beyond that point.

Team policing, being an organizational change, would likely experience some start-up problems (e.g., officers getting to know their assigned neighborhoods); but when it ended, there would be little carryover to the next segment of the time series.
It is clear from Figure 3.9 that when team policing and neighborhood watch operated together (period 3), the level of burglaries was lowest in the time series. When team policing operated alone (period 4), burglaries increased somewhat, but were still substantially lower than they were for either of the periods (2 and 5) when neighborhood watch operated alone.

Based on Figure 3.9 and the findings from evaluating neighborhood watch at the block and area of the city levels, it is reasonable (although not definitive) to conclude that the team-policing program was primarily responsible for reducing burglaries. Our conclusion is not definitive—very few program evaluation findings are—but it is consistent with the evidence and serves to reduce the uncertainty around the question of relative program effectiveness.

The evaluation of the York crime prevention programs employed several different research designs. Time-series designs can be useful for assessing program outcomes in situations where data exist for key program logic constructs before and after (or during) program implementation. The implicit design used to survey the block watch captains is perhaps the most vulnerable to internal validity problems of any of the possible research designs evaluators can use. For implicit designs, modeled as \( (X \ O) \), there is neither a pretest nor a control group, so we may not be able to reduce any of the uncertainty around the question of whether the program caused the observed outcomes.

But even in this situation (which is actually quite common for program evaluators), there are ways of strengthening your hand. Suppose you have been asked to evaluate a program that offers small businesses subsidies to hire people aged 17 to 24. The objectives of the program are to provide work experience, improved knowledge of business environments, and encouragement to either start their own business or pursue business-related postsecondary education.

As a program evaluator, it would be worthwhile having a **comparison group** who did not get the program, so that constructs like “increased knowledge of business practices” could be measured and the results compared. But that may not be possible, given resource constraints. Instead, you might still be expected to evaluate the program for its effectiveness and be expected to do so by focusing on the program alone.

One way to reduce uncertainty in the conclusions drawn is to acknowledge the limitations of an implicit \((X \ O)\) design, but apply the design to different stakeholder groups. In the business experience program evaluation, it would make sense to survey (or interview) a sample of clients, a sample of employers, and the program providers. These three viewpoints on the program are complementary and allow the evaluator to triangulate the
perspectives of stakeholders. In effect, the X O research design has been repeated for three different variables: client perceptions, employer perceptions, and program provider perceptions.

**Triangulation** is an idea that had its origins in the literature on measurement. As a measurement strategy, triangulation is intended to strengthen confidence in the validity of measures used in social research.

Once a proposition has been confirmed by two or more independent measurement processes, the uncertainty of its interpretation is greatly reduced. The most persuasive evidence comes through a triangulation of measurement processes. If a proposition can survive the onslaught of a series of imperfect measures, with all their irrelevant error, confidence should be placed in it. (Webb, 1966, p. 3)

In our situation, triangulation that is focused on the question of whether the program was effective can at least establish whether there is a concurrence of viewpoints on this question, as well as other related issues. It does not offer a firm solution to the problem of our vulnerable research design, but it offers a workable strategy for increasing confidence in evaluation findings.

**Construct Validity**

Figure 3.5 indicates that construct validity is about being able to generalize from the variables and their observed relationships in a program evaluation back to the constructs and their relationships in the program logic. **Measurement** is about translating constructs into observables, and that at the “level” of observables, we work with variables, not the constructs themselves. In the evaluation of the York Neighborhood Watch Program, for example, the key dependent variable was the monthly total of burglaries that had been reported (and recorded) by the police department.

The evaluators wanted to learn whether the program was effective in reducing burglaries in York. The construct implicit in this question is the number of burglaries committed—we have measured the construct by assuming that reported burglaries is a valid measure of burglaries committed.

Basically, construct validity problems can result from the way the program has been operationalized in the evaluation, or the ways that outputs, linking constructs, or outcomes have been measured.

An example of a construct validity problem that involves the way the program has been operationalized can be taken from the Perry Preschool Program evaluation (Berrueta-Clement, 1984; Schweinhart, Barnes, & Weikart, 1993).
In the Perry Preschool experiment (1962–1967), children from families in Ypsilanti, Michigan, aged 3 and 4, were randomly assigned in cohorts to program and no-program groups. The purpose of the experiment was to see what the effects of preschool are on a set of variables related to academic achievement and social adjustment.

The authors reported that younger siblings of children who had been assigned to the program were included in the program when they reached preschool age (Schweinhart et al., 1993). The program included a classroom preschool component as well as weekly home visits. But for families with two or more preschool program children, it is possible that siblings reinforced each other and created a different at-home environment. Under this condition, does the program include the two intended components, or those plus sibling reinforcement? This ambiguity is a construct validity problem for the evaluation.

In an evaluation of a server training program, the evaluators assigned matched pairs of drinking establishments in Thunder Bay, Ontario, to program and no-program conditions (Gliksman, McKenzie, Single, et al., 1993). Managers in the establishments had been asked if they were willing to have their servers trained, but were cautioned not to tell their servers about the evaluation. Given the incentives for managers to “look good” or “cooperate,” it is possible that some managers mentioned the evaluation to their servers. The construct validity problem created is: What is the program—server training, or server training plus the informal influence of bar managers?

Shadish, Cook, and Campbell (2002) point out that construct validity issues and internal validity issues are both about factors that confound our interpretation of the results we observe in an evaluation. Internal validity threats are sets of factors that could explain the outcomes we observe, regardless of whether the program/treatment happened. If a seatbelt law is implemented in a state at the same time that a program is implemented to crack down on speeding, a reduction in traffic fatalities could be due to the seatbelt law, regardless of whether the program was implemented.

Construct validity issues arise from the way the treatment/program was implemented. They cannot occur if there has not been a treatment or program in place. In the example from the Perry Preschool experiment mentioned above, the construct validity problems are created by the possibility that older siblings of preschool group children, having been to preschool themselves, act as home tutors and socializers for their younger siblings. This situation results in ambiguity in just what the program is: preschool (including home visits) or preschool plus sibling influence.

In an earlier formulation of threats to internal validity (Cook & Campbell, 1979), four threats were discussed that focused on the context for a
They could not have occurred if the treatment/program had not been implemented. Because they affect our ability to generalize from the observed findings back to the constructs in the program theory, they are included here as examples of construct validity threats.

1. **Diffusion of treatments**: People can sometimes communicate about their program experiences to members of the control group.

   Example: Two groups of employees in a company are selected to participate in a team-building experiment. One group participates in team-building workshops. The other group (who may have an opportunity to take the workshop later) serves as the control group. Employees communicate, and some of the skills are transferred informally.

2. **Compensatory equalization of treatments**: The group that is not supposed to get the program is offered the program to avoid criticisms of withholding something that could be beneficial.

   Example: One neighborhood/area gets a neighborhood watch program, other areas (that were control neighborhoods) insist on the right to the program in their areas so the program spreads.

3. **Compensatory rivalry**: The no-program group is given a “better” version of the existing program, which tends to cancel out the differences between the new program and the existing program.

   Example: Two versions of a course are offered in an undergrad program, one uses interactive multimedia tools while the other is taught in a conventional classroom format. The classroom instructor works twice as hard as she usually does, not to be outdone by this new instructional method.

4. **Resentful demoralization**: The control group perceives unfair treatment and reacts negatively.

   Example: Those persons not getting a program to test the effects of class size on learning (halving the size of classes) complain to the instructor and demand equal treatment. The administration refuses and students threaten to not take any of the remaining tests in the course.

Cook and Campbell (1979) and Shadish, Cook, and Campbell (2002) present lists of circumstances that might weaken construct validity in an
Evaluation. In summarizing the ways of minimizing this set of problems, they suggest that evaluators need to: make sure that constructs are clearly defined so that they can be measured appropriately; make sure that constructs are differentiated so that they do not overlap as measures are developed; develop "good" measures, that is, measures that produce unbiased information; and, ideally, use multiple measures of key constructs.

Shadish, Cook, and Campbell (2002) also point out that construct validity threats can be caused by the fact that people know they are a part of an evaluation process. Participant expectations can influence behaviors, confounding attempts to generalize the actual findings back to the constructs or program theory. One example of the way that participant behavior can confound an experiment is the Hawthorne effect.

In a worker-productivity experiment in the 1930s in the U.S., the experimenters discovered that being part of an experiment produced an effect regardless of the levels of the experimental variables being manipulated (Roethlisberger, Dickson, & Wright, 1939). No matter what conditions the experimenters varied (e.g., lighting level, speed of the assembly line, variability of the work), the results indicated that any manipulation increased productivity. This outcome is usually referred to as a Hawthorne effect, named after the location where the research occurred. Construct validity was compromised by the behavior of the workers.

External Validity

Figure 3.5 suggests that external validity builds on the "validities" that have been discussed thus far: statistical conclusions, internal, and construct validity. External validity is about generalizing the causal results of a program evaluation to other settings, other people, other program variations, and other times.

Fundamentally, we want to know whether the program, the evaluation results, and the clients/participants and the setting(s) are representative of other circumstances in which we might wish to apply the program or the evaluation results. In the evaluation of the server-training program in Thunder Bay, Ontario, how confident are we that the results of that evaluation are generalizable, given that the drinking establishments were chosen for convenience by the evaluation team. Are the results unique to that setting, or would they be representative of what would happen if server training programs were implemented in other establishments in Thunder Bay, or other cities in the province of Ontario?

Shadish, Cook, and Campbell (2002) suggest five categories of external validity threats. In each one, the causal results obtained from a given evaluation
are threatened by factors that somehow make the results unique. Four of the
five threats are interaction effects that reflect their concern with generalizing
to other units of analysis (typically people), other program variations, other
outcome variations, and other settings.

1. Interaction between the causal results of a program and the
people/participants.

Example: The Perry Preschool experiment was conducted with a
mix of boys and girls, but the program appeared to work much
more effectively for girls (Schweinhart et al., 1993). Does that
mean that efforts to implement the program elsewhere would
encounter the same limitation? Why did the program not pro-
duce the same results for boys?

2. Interaction between the causal results of a program and the
treatment variations.

Example: The Perry Preschool experiment had two main com-
ponents. Children in the program cohorts attended preschool 5 days
a week during the school year for 2 years. In addition, there were
home visits to the families each week to reinforce the school set-
ting and establish acceptance of preschool in the family. This ver-
sion of the program is expensive. Because it is called a “preschool
program” other implementations of the program might skip the
home visits. Would that undermine the program results?

3. Interaction between the causal results of a program and outcome
variations.

Example: A state operated program that is intended to train
unemployed workers for entry-level jobs succeeds in finding job
placements (at least 6 months long) for 60% of its graduates. A
comparison with workers who were eligible for the program but
could not enroll due to space limitations, suggests that the pro-
gram boosted employment rates from 30% to 60%—an incre-
mental effect of 30%. Another state is interested in the program
but wants to emphasize long-term employment (2 years or
more). Would the program results hold up if the definition of the
key outcome were changed?

4. Interaction between the causal results of a program and the setting.

Example: The Abecedarian Project (F. A. Campbell & Ramey,
1994) was a randomized experiment intended to improve the
school-related and cognitive skills of children for poor families in a North Carolina community. The setting was a university town, where most families enjoyed good incomes. The project was focused on a segment of the population that was small, relative to the rest of the community. The program succeeded in improving cognitive, academic, and language skills. But could these results, robust though they are, be generalized to other settings where poor, predominately minority families resided?

There is a fifth threat to external validity that also limits generalizability of the causal results of an evaluation. Shadish, Cook, and Campbell (2002) call this “context-dependent mediation.” An example would be a situation where a successful crime prevention program in a community used existing neighborhood associations to solicit interest in organizing blocks as neighborhood watch units. Because the neighborhood associations were well established and well known, the start-up time for the crime prevention program was negligible. Members of the executives of the associations volunteered to be the first block captains, and the program was able to show substantial numbers of blocks organized within 6 months of its inception.

The program success might have been mediated by the neighborhood associations—their absence in other communities could affect the number of blocks organized, and the overall success of the program.

**TESTING THE CAUSAL LINKAGES IN PROGRAM LOGIC MODELS**

Research designs are intended as tools to examine causal relationships. In program evaluation, there has been a general tendency to focus research designs on the linkage between the program as a whole and the observed outcomes. Although we learn by doing this, the current emphasis on elaborating program descriptions as logic models presents situations where our logic models are generally richer and more complex than our research designs. When we evaluate programs, we generally want to examine the linkages in the logic model, so that we can see whether (for example) levels of outputs are correlated with levels of linking constructs, and they, in turn, are correlated with levels of outcomes. Research designs are important in helping us to isolate each linkage so that we can assess whether the intended causal relationships are corroborated. But as we said earlier, designing an evaluation so that each program component can successively be isolated to rule out rival hypotheses, is expensive and generally not practical.
Having discussed some of the findings from the York crime prevention program, we can display the program logic in Table 3.3.

The intended outcome is a reduction in burglaries committed. To achieve that objective, the program logic specifies a series of steps, beginning with organizing city blocks into neighborhood watch blocks. If the program logic works as intended, then our program theory will have been corroborated in this implementation.

We can also summarize the research designs, the units of analysis and the key comparisons that were included in the York crime prevention program. Table 3.4 displays these features of the program.

By reviewing Figure 3.9, we can see that there were four different research designs in the evaluation and that none of them facilitated an examination of the whole logic model. Each design focused on one construct in the model, and the data collected permitted the evaluators to see how that part of the logic model behaved.

As an example, the block captain interviews focused on attendance at meetings, which is an important linking construct in the logic model. But they did not measure increased awareness of prevention techniques,

<table>
<thead>
<tr>
<th>Components</th>
<th>Implementation Objectives</th>
<th>Outputs</th>
<th>Linking Constructs</th>
<th>Intended Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighborhood watch blocks</td>
<td>To organize neighborhood watch blocks in the target areas of the city</td>
<td>Number of blocks organized</td>
<td>Attendance at block meetings</td>
<td>Reduced in burglaries committed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of block meetings held</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Increased awareness of burglary prevention techniques</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Increased application of prevention techniques</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Improved home security</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3 Program Logic for the York Crime Prevention Program
increased applications of techniques, or improved home security. In fact, those three linking constructs were not measured in the evaluation. To do so would have required a survey of neighborhood residents, and there were insufficient resources to do that. Likewise, the time-series designs facilitated tracking changes in blocks that were organized, and burglaries were reported over time but were not set up to measure other constructs in the program logic.

In sum, each design addresses a part of the program logic, and helps us to see if those parts are behaving as the logic intended. But what is missing is a way to test the connections between constructs. Even if blocks are organized and neighbors attend block watch meetings, we do not know whether the steps leading to reduced burglaries have worked as intended.

Why not design program evaluations so that the whole logic model is tested? The main reason is cost. In the York crime prevention evaluation, a full test of the program logic in Table 3.3 (testing the intended linkages among outputs, linking constructs, and outcomes) would require that all the constructs be measured using the same unit of analysis.

Suppose that random samples of residents in the target and control neighborhoods were enlisted to participate in a 4-year study of crime prevention effectiveness in the city. Initially (2 years before the program started), each home (the unit of analysis) would be surveyed to find out if any had experienced burglaries in the past 12 months; householders would also be asked about their participation in any crime prevention activities, their

<table>
<thead>
<tr>
<th>Research Design</th>
<th>Unit of Analysis</th>
<th>Key Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single time series</td>
<td>Time (monthly)</td>
<td>The number of blocks organized (an output variable) was recorded monthly by the police department and compared over time</td>
</tr>
<tr>
<td>Implicit design</td>
<td>Block captains</td>
<td>Block captains were interviewed at one point in time to obtain estimates of the numbers of meetings held (output), and resident attendance (linking construct)</td>
</tr>
<tr>
<td>Comparative time series</td>
<td>Time (monthly)</td>
<td>Reported burglaries (an outcome) were compared on a monthly basis in the two areas of York that were “program” and “control”</td>
</tr>
<tr>
<td>Single time series</td>
<td>Time (monthly)</td>
<td>Reported burglaries (an outcome) were compared each month city-wide before and after the program was implemented</td>
</tr>
</tbody>
</table>
awareness of burglary prevention techniques, and their existing home security measures.

This survey could be repeated (using the same sample) each year for 4 years (2 before the program, and 2 after implementation). After the program was implemented, the survey participants would also be asked whether they participated in the program, and if they did, how frequently they attended block watch meetings, whether the block watch meetings increased their awareness (they could be "tested" for their level of awareness), whether they were taking any new precautions to prevent burglaries, and finally, whether their homes had been burglarized in the previous 12 months.

Notice that this information “covers” the linkages in the logic model, and by comparing responses between the target and control neighborhoods, and responses within households over time, we could assess the causal linkages in all parts of the model. For example, comparisons between program and no-program residents after the program was implemented would indicate whether program residents were more likely to be aware of burglary prevention methods, more likely to apply such methods, and more likely to have more secure homes.

The Perry Preschool experimental study of the long-term effects of preschool on a sample of Black children in Ypsilanti, Michigan, has been able to follow the program and no-program children from age 3 to age 27 (Schweinhart et al., 1993). Over that time, multiple questionnaires were administered to participants and data gathered on a large number of separate measures of school achievement, marital and family relations, criminal behaviors, employment, earnings, and other personal, social, and economic variables.

The researchers have constructed and tested a logic model for the experiment. Key outcomes as of age 27 include: monthly earnings and lifetime arrests. These outcomes are explained in the model by years of schooling, school motivation in kindergarten and primary grades, IQ after 1 preschool year, whether the person had preschool or not, and finally, family characteristics at age 3 and 4 (Schweinhart et al., 1993).

Given sufficient money, time, and control over the program situation, it is possible to fully test program logics, as the Perry Preschool experiment demonstrates. Testing the program theory, using approaches that permit full tests of the logic model, is an important and growing part of the field (Bickman, 1987; Rogers, Hacsi, Petrosino, & Huebner, 2000). An additional example of this strategy is the evaluation of the WINGS project (Hennessy & Greenberg, 1999). The WINGS (Women In Group Support) project was focused on reducing sexual risk-taking among vulnerable women in the United States. The project was designed as a randomized experiment and
implemented in multiple sites in 1992. The evaluators set up a logic model of the intended program theory that could be tested as a recursive causal model. In the program, women in the program groups were a part of 6 weekly sessions that provided information, modeled behaviors (including asking partners to use condoms), and provided opportunities for the participants to role play situations where the issue of safe sex would happen.

The evaluation demonstrated that the program was effective in reducing the infection rates from sexually transmitted diseases. The logic model was fully tested, to determine whether the intended linkages among constructs were corroborated by the data.

Donaldson’s and Gooler’s (2002) evaluation of the California Wellness Foundation’s Work and Health Initiative introduces the role of evaluator judgment into the technical process of testing program theory. In an interview by Jody Fitzpatrick for the American Journal of Evaluation, Stewart Donaldson describes his approach to theory-driven program evaluation, based on two decades of experience (Fitzpatrick, 2002). Although there is clearly an important technical aspect to testing program theories, Donaldson implies that his own experience and judgment are also important parts of this process.

Program theory clearly represents the presumed active ingredients of the program and how they lead to the desired outcomes. You’re trying to consider whether the paths leading to this outcome have any chance at all for accounting for a significant portion of the variance in the desired outcome . . . I take each path and really think carefully about the timing and when it should occur and whether there are key moderators of that path. (p. 353)

When we look at the practice of evaluation, rarely do evaluators have the time, the resources, and the control to proceed as has been suggested above. Instead, we are usually expected to conduct program evaluations after the program has been implemented, using (mostly) existing data. These constraints usually mean that we can examine parts of program logics with evidence and other parts with our own observations and professional judgments.

The discussion so far has emphasized the connections between research designs and the central question of determining program effectiveness in program evaluations. Research design considerations need to be kept in
Performance measurement parallels key parts of what we covered in Chapter 2. Identifying the key constructs that describe a program and outlining its intended cause and effect linkages are central to any effort to build a viable performance measurement system (Mayne, 2001).

Program evaluations are usually discrete projects: they have a beginning, middle, and an end point. Often, ad hoc committees of stakeholders are struck to oversee the process, and usually, the client(s) are involved in the evaluation from its first steps. Typically, program evaluations are commissioned because there are questions that need to be answered that cannot be addressed with existing administrative data sources and management information systems.

Performance measurement systems are often put together to serve the purposes of improving the efficiency and effectiveness of programs and accounting for program results to stakeholders (agency executives, elected officials, and legislatures) (Hatry, 1999). When performance measures are developed and implemented, it is usually with the intention of providing ongoing information on the outputs and outcomes of programs.

A key issue for any of us who are interested in developing and implementing credible performance measures is the expectation, on the one hand, that the measures we come up with will tell us (and other stakeholders) how well the observed outcomes approximate the intended program objectives, and on the other hand, constructing measures that actually tell us what the program has accomplished. This latter concern is, of course, our incrementality question: What differences did the program actually make? Answering it entails wrestling with the question of the extent, if any, to which the program caused the observed outcomes.

Performance measurement systems typically are not well equipped to tell stakeholders whether the observed outcomes were actually caused by the program. They can describe the observed outcomes, and they can tell us whether the observed outcomes are consistent with program objectives, but there is usually a shortage of information that would get at the question of whether the observed outcomes were the result of program activities (Newcomer, 1997).

If we think of performance measurement as a process of tracking programs over time, we can see that many of the measures built into such systems are, in fact, time series. Variables are measured at regular intervals and the changes in their levels are assessed. Often, trends and levels of performance variables are compared to targets or benchmarks. In some situations, where a change in program structure or activities has been
implemented, it is possible to track the before-after differences, and at least see whether the observed changes in levels and trends are consistent with the intended effects.

In situations where we would want to use time-series data to look for effects that are consistent (or inconsistent) with intended outcomes, we need continuity in the way variables are measured. If we change the way the measures are taken, or change the definition of the measure itself (to improve its relevancy for current program and policy priorities), we may jeopardize its usefulness as a way to assess cause and effect linkages.

As we have seen in Chapter 2, logic models have two purposes: they describe a program (with components, elements, and implementation objectives), and they delineate intended cause and effect linkages (outputs, linking constructs, and outcomes). Outputs are typically thought of as measures of the work done, and we have conceptualized them as being part of the program, as distinct from being in its environment.

In program evaluations and in performance measurement systems, outputs are typically viewed as attributable to the program—one does not usually need an elaborate research design to test whether the outputs were caused by the program. That means that performance measures that focus on outputs typically can claim that they are measuring what the program actually produced.

Since outputs are generally more likely to be under the control of program managers, there is normally a much greater willingness to report outputs, as opposed to outcomes, as performance measures. As stakeholders insist that performance measures should focus on outcomes that are further along in the logic chain, there is typically much less willingness to “own” the results, that is, accept them as being due to program activities alone.

Thus, using performance measures as a way to determine what a program actually accomplished presumes that the causality issue has been settled. For outputs, that is less an issue than it is for outcomes. As we see in Chapter 10 the incentives for managers play a key role in how willing they are to participate in developing and using performance measures as ways of assessing what their programs accomplished.

**SUMMARY**

This chapter focuses on the issue of how research designs can support program evaluators who want to assess whether a program has achieved its objectives. Examining whether the program was effective is a key question in most evaluations, regardless of whether they are formative or summative.
Although randomized experiments are often beyond the resources and control required to design and implement them well, the logic of experimental designs is important to understand if evaluators want to address questions of whether the program caused the observed outcomes. The three conditions for establishing causality—temporal asymmetry, co-variation between the causal variable and the effect variable, and no plausible rival hypotheses—are at the core of all experimental designs.

Through randomized experimental designs, whereby members are assigned randomly to either the program or the control group, it is possible to assert that the two groups are equal in all respects before the program begins. When the program is implemented, the only difference between the two groups is the program itself. This makes it possible to isolate the incremental effects of the program on the participants. Typically, in randomized experiments, we say that we have controlled for threats to the internal validity of the research design, although there can be internal validity problems with the implementation of experiments (Olds et al., 2000).

In assessing research designs, we should keep in mind that the four different kinds of validity are cumulative. Statistical conclusions validity is about determining whether the program and the outcome variable(s) co-vary. Co-variation is a necessary condition for causality. Internal validity builds on statistical conclusions validity and examines whether there are any plausible rival hypotheses that could explain the observed co-variation between the program and the outcome variable(s). Ruling out plausible rival hypotheses is also a necessary condition for causality. Construct validity generalizes the data-based findings back to the constructs and intended linkages in the logic model. Finally, external validity generalizes the results of the program evaluation to other times and places.

Departures from randomized experimental designs can also work well for determining whether a program caused the observed outcomes. One of the most common quasi-experimental designs is the time series, where a program is implemented part way through the time series. Single and multiple time series make it possible to selectively address internal validity threats. Deciding whether a particular threat to internal validity is plausible or not, entails using professional judgment, as well as evidence from the evaluation itself.

When we develop a logic model of a program, we are specifying a theory of the program. Ideally, we want to test this theory in an evaluation. Most program evaluations do not permit this because the resources are not there to do so. Further, most evaluations use several different research designs, each having the capability of testing a part of the logic model, not all of it collectively.

Theory-driven program evaluations are designed so that it is possible to test the full logic model. By specifying a single unit of analysis that permits
data collection for all the constructs in the model, statistical methods can be used to test each linkage in the model, controlling for the influences on that link, of other paths in the model. Although this approach to program evaluations typically requires extensive resources and control over the evaluation process, it is growing in importance as we realize that linking logic models to research designs that facilitate causal tests of linkages in the models is a powerful way to assess program effectiveness.

**DISCUSSION QUESTIONS**

1. The following diagram shows several weak research designs that have been used in an evaluation. The “O” variable is the same for the entire diagram and is measured in such a way that it is possible to calculate an average score for each measurement. Thus, O₁, O₂, O₃, O₄, and O₅ all represent the same variable and the numbers in brackets above each represent the average score for persons who are measured at that point. All the persons in group 1 are posttested; group 2 had been randomly divided into two subgroups and one subgroup has been pretested and the other one has not. Notice that all the members in group 2 got the program. Finally, for group 3, there is a pretest only.

- Examine the averages that correspond to the five measurements and decide which threat to internal validity of the overall research design is clearly illustrated. Assume that attrition is not a problem, that is, all persons pretested are also posttested. Explain your answer, using information from Table 3A.1.

<table>
<thead>
<tr>
<th>Table 3A.1</th>
<th>What Threat to Internal Validity Is Illustrated by This Patched-up Research Design?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group 1</strong></td>
<td>(6.0)</td>
</tr>
<tr>
<td>X</td>
<td>O₁</td>
</tr>
<tr>
<td><strong>Group 2</strong></td>
<td>(4.0)</td>
</tr>
<tr>
<td>R</td>
<td>O₂ X (7.0)</td>
</tr>
<tr>
<td>R</td>
<td>X (6.0)</td>
</tr>
<tr>
<td></td>
<td>O₄</td>
</tr>
<tr>
<td><strong>Group 3</strong></td>
<td>(4.0)</td>
</tr>
<tr>
<td></td>
<td>O₅</td>
</tr>
</tbody>
</table>
Table 3A.2  Statistical Tools for Program Evaluation

<table>
<thead>
<tr>
<th>Levels of Measurement</th>
<th>Describing One Variable</th>
<th>Generalizing One Variable</th>
<th>Describing Associations Between Two Variables</th>
<th>Generalizing Associations Between Two Variables</th>
<th>Describing Associations Among Three Variables</th>
<th>Generalizing Associations Among Three Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal or nominal by nominal</td>
<td>• Mode</td>
<td>• Variation ratio</td>
<td>• Phi</td>
<td>• Chi square (null hypothesis of no association between the variables)</td>
<td>• Elaborating contingency tables and comparing values of phi or Cramer’s V</td>
<td>• Pooled chi square</td>
</tr>
<tr>
<td>Ordinal or ordinal by ordinal</td>
<td>• Median</td>
<td>• Semi-interquartile range</td>
<td>• Tau B</td>
<td>• Significance test (null hypothesis of no association between the variables)</td>
<td>• Elaborating contingency tables and comparing values of tau B or tau C or gamma</td>
<td></td>
</tr>
<tr>
<td>Interval or interval by interval</td>
<td>• Mean</td>
<td>• Standard deviation</td>
<td>• Confidence intervals</td>
<td>• Pearson’s r</td>
<td>• Significance test for Pearson’s r (null hypothesis of no slope on the regression line, i.e., b = 0)</td>
<td>Multiple regression analysis:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Single variable hypothesis tests</td>
<td></td>
<td></td>
<td>• Multiple R</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Unstandardized regression coefficients (b coefficients)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Standardized regression coefficients (beta coefficients)</td>
</tr>
<tr>
<td>Nominal by ordinal</td>
<td>• Phi</td>
<td>• Chi square (Ho: no association)</td>
<td>• Elaborating contingency tables and comparing values of phi or Cramer’s V</td>
<td>• Elaborating contingency tables and comparing values of phi or Cramer’s V</td>
<td>• Pooled chi square</td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
Table 3A.2 (Continued)

<table>
<thead>
<tr>
<th>Levels of Measurement</th>
<th>Describing One Variable</th>
<th>Generalizing One Variable</th>
<th>Describing Associations Between Two Variables</th>
<th>Generalizing Associations Between Two Variables</th>
<th>Describing Associations Among Three Variables</th>
<th>Generalizing Associations Among Three Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordinal by interval</td>
<td></td>
<td></td>
<td>• Tau B or tau C or gamma or Spearman’s rho for cross tabs</td>
<td>• Significance test (for cross-tabulations)</td>
<td>• Tau B or tau C or gamma for cross-tabs</td>
<td>Two-way ANOVA: Overall F test</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• One-way ANOVA: Eta (for situations where the independent variable is ordinal)</td>
<td>• One-way ANOVA: F test (for situations where the independent variable is ordinal)</td>
<td>• Two-way ANOVA: partial Eta’s and overall Eta (where the two independent variables are ordinal)</td>
<td>• Partial F tests for each independent variable</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Phi or Cramer’s V or contingency coefficient for cross tabs</td>
<td>• Chi square (for cross-tabs)</td>
<td>• Multiple regression analysis (where the dependent variable is interval)—same statistics as interval by interval (above)</td>
<td>• Tests for interactions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• One-way ANOVA: Eta where the independent variable is nominal</td>
<td>• One-way ANOVA: F test (for situations where the independent variable is ordinal)</td>
<td>• Two-sample t test (where the interval variable is dependent)</td>
<td>Multiple regression—same statistics as ordinal by interval</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Pearson’s r</td>
<td>• Two-sample t test (where the interval variable is dependent)</td>
<td>• Multiple regression (where the interval variable is dependent)—same statistics as above</td>
<td>Two-way ANOVA—same statistics as ordinal by interval</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• significance test for Pearson’s r</td>
<td>• Multiple regression (where the independent variable is dependent)—same statistics as above</td>
<td>• Multiple regression—same statistics as interval by interval</td>
<td>• Multiple regression—same statistics as interval by interval</td>
</tr>
</tbody>
</table>
2. What is a key difference between internal validity and external validity in research designs?

3. What is the difference between testing and instrumentation as threats to the internal validity of research designs?

4. What is the difference between history and selection as threats to the internal validity of research designs?

5. A nonprofit organization in a western state has operated a 40-hour motorcycle safety program for the past 10 years. The program permits unlicensed, novice motorcycle riders to learn skills that are believed necessary to reduce accidents involving motorcyclists. Upon completing the 1-week course, trainees are given a standard state drivers test for motorcycle riders. If they pass, they are licensed to ride a motorcycle in the state. The program operates in one city and the training program graduates about 400 motorcyclists per year. The objective of the program is to reduce the number of motor vehicle accidents involving motorcyclists. Because the program has been targeted in one city, the effects would tend to be focused on that community. The key question is whether this course does reduce motorcycle accidents for those who are trained. Your task is to design an evaluation that will tell us whether the training program is effective in reducing motorcycle accidents. In designing your evaluation, pay attention to the internal and construct validities of the design. What comparisons would you want to build into your design? What would you want to measure to see whether the program was effective? How would you know if the program was successful?

REFERENCES


