

---

# **Program Evaluation: A Practical Guide to Discovering What Works**

**Howard J. Shaffer**

**Matthew N. Hall**

**Joni Vander Bult**

*Harvard Medical School, Division on Addictions  
Addiction Technology Transfer Center of New England*

---

Addiction Technology Transfer Center of New England Technical Report # EV-122297  
Providence, RI  
December 22, 1997

# Program Evaluation: A Practical Guide to Discovering What Works

## Table of Contents

<b>Introduction</b>	<b>3</b>
<b>Identifying the Issues</b>	<b>3</b>
<i>Why Conduct an Evaluation?</i>	3
<b>Research Methods: Potential Problems in Evaluation Projects</b>	<b>4</b>
<i>Research Terminology</i>	5
<i>Threats to Validity</i>	6
<i>Threats to Generalizability</i>	9
<b>Alternative Study Designs: How to Avoid Research Problems</b>	<b>10</b>
<i>The “Pre-Experimental” Designs</i>	10
Design #1: The “One-Shot” Case Study	10
Design #2: The One-Group Pretest/Posttest Design	11
Design #3: The Static Group Comparison	11
<i>The Experimental Designs</i>	12
Design #4: The Pretest/Posttest Control Group Design	12
Design #5: The Solomon Four-Group Design	12
Design #6: The Posttest-Only Control Group Design	13
<b>Preparing to Conduct Research: Steps to Planning and Implementing Evaluation Projects</b>	<b>13</b>
<i>Determining the Research Questions</i>	14
<i>Developing the Research Protocol</i>	15
Building an Evaluation Instrument	15
Identifying the study population	17
Identifying the Appropriate Research Design	17
Power Analysis	18
Determining the Sampling Methods	19
Response Rates	21
Planning Data Coding and Entry	22
Planning Data Analysis	22
Planning for Data Interpretation	23
<i>Pilot Testing and Revising the Protocol</i>	23
<i>Implementing the Research Program</i>	23
<i>Drawing Conclusions</i>	24
<b>Making Program Decisions: Using Evaluation Research</b>	<b>24</b>
<b>References</b>	<b>26</b>

# Program Evaluation: A Practical Guide to Discovering What Works

## Introduction

This guide is intended primarily for school administrators, community leaders, and others who do not have a background in evaluation research but are interested in evaluating programs (e.g., prevention, curriculum, or other organized activities intended to yield Specific outcomes). This guide is also intended for those who simply want a quick and handy resource for the important elements of program evaluation. For example, school superintendents often want to implement new substance abuse prevention programs in the middle schools of their towns. How will they know if these programs are effective? How will they know if a program inadvertently makes a situation worse? This guide aims to clarify the conceptual issues associated with research questions of this sort, present alternative approaches to evaluation, and identify the pros and cons associated with each of these various approaches.

The purpose of this document is to provide practical guidelines for developing and implementing evaluations of a wide range of applied programs. Throughout this report, we will use school-based substance abuse prevention and education programs as the primary examples of the types of programs that could be examined and improved through evaluation. However, readers should keep in mind that there is a broad range of programs that could be evaluated, including other school curricula, community-based prevention programs, public awareness campaigns, treatment programs, and many other types of activities.

## Identifying the Issues

### *Why Conduct an Evaluation?*

The first principle of medical ethics is to do no harm. This maxim exists because the best of intentions may unwittingly lead to adverse consequences. The need for program evaluation is dictated by a simple premise: unless a program is evaluated, we do not know whether it is producing positive, neutral, or negative results. It is easy to assume that a program's outcome will be obvious and straightforward—for example, that a middle school substance abuse prevention program will prevent substance abuse among middle school students. Unfortunately, reality seldom conforms to these assumptions in a straightforward manner. In fact, substance abuse prevention programs can a) have no effect; b) change knowledge about substance abuse but not drug using behaviors; c) decrease substance use as planned; d) inadvertently increase substance use; or e) have a range of other outcomes. To the surprise of administrators, too often evaluation research reveals that programs have outcomes quite different from those that program developers intended. For example, some researchers have suggested that informational approaches to substance abuse prevention (e.g., providing information on the potential dangers of substance abuse) may actually increase substance use among adolescents by stimulating their curiosity (Botvin & Botvin, 1992).

Before undertaking an evaluation of any program, the first step is to identify with precision what is the specific purpose of the program under consideration. For example, the purpose of a middle school substance abuse prevention curriculum might initially seem to be fairly straight forward: that is, to keep middle school students from using psychoactive drugs. However, a program of this type actually could have a number of different outcomes as its goal. For example, this prevention curriculum could have any one or a combination of several of the following goals: 1) preventing substance use from starting; 2) preventing harms caused by substance use from starting; 3) limiting harms that have started; 4) strengthening factors that limit substance use (e.g., improving family relationships, strengthening interests in activities that are incompatible with substance use); or, 5) a range of other potential outcomes. Program evaluators assessing this program would measure variables related to its specific objectives. Clearly, if program planners are unclear about the specific goal or goals of a program, an evaluation would yield little useful or meaningful information.

Evaluation research efforts focus on determining the ability of the program to reach its stated objectives (e.g., Guttentag & Struening, 1975; Struening, & Guttentag, 1975). However, prior to developing an evaluation project, investigators and program staff might find it productive and interesting also to ask themselves “why are these objectives important?” Valuable program evaluation studies emerge by choosing to study very specific objectives that are also important. The next section of this guide describes the issues evaluators should consider when planning their evaluations and the corresponding research methods that will yield the information they want.

## **Research Methods: Potential Problems in Evaluation Projects**

In all types of research projects, investigators strive to achieve a set of ideal analytic circumstances known as “experimental conditions.” When investigators implement a research project that allows them to manipulate the variable of primary interest and keep other factors constant, an “experimental” design exists. For example, in a study of the effects of a new blood pressure medication, the variable of interest would be the new medication; the experimenters would manipulate this variable by giving it to one group of subjects and not giving it to another group. There would be many “other factors” that the experimenters would want to keep equivalent for the two groups. For example, the two groups should be fairly comparable in terms of age, gender, diet, use of other medications, degree of obesity, smoking, and a range of other potentially influential variables. If investigators did not control these “other factors,” any differences identified between the two groups at the end of the study could be the result of one of these other factors (e.g., a higher percentage of smokers in one group).

As this example illustrates, it is necessary to compare the group (or groups) that experience the program of interest to a comparison group (or groups) that do not experience the program but are similar in other ways. In addition, there is a range of other factors that should be addressed to achieve the proper experimental conditions in a program evaluation. Therefore, before considering the variety of research designs that can accomplish these evaluative tasks, and the relative strengths

and weaknesses of each approach, we will describe the terminology associated with research designs and some of the threats to the validity and generalizability of evaluation research efforts.

### ***Research Terminology***

In discussing research methodology, it is helpful to use standardized terminology to present concepts in a consistent and understandable way. In this report, we will refer to the program being evaluated as the “experimental condition” and the group of people who experience the program as the “experimental group.” Similarly, we will refer to the comparison condition as the “control condition” and the comparison group as the “control group.” In addition, we will use the following shorthand, adapted from Campbell & Stanley (1963), to refer to specific aspects of a research design:

**X** - exposure of a group to the experimental condition (i.e., the program being evaluated)  
**O** - the process of observation or measurement (e.g., a pretest, a posttest)  
**R** - the process of randomly assigning respondents to groups

In the tables that use this shorthand to describe different research designs, we will follow a few other conventions: left to right indicates the passage of time, and symbols in a vertical column occur simultaneously. The following table presents an example of these customs:

Sample Table				
	Time 1	Time 2	Time 3	Time 4
<b>Group 1</b>	<b>R</b>	<b>O<sub>1</sub></b>	<b>X</b>	<b>O<sub>2</sub></b>
<b>Group 2</b>	<b>R</b>	<b>O<sub>3</sub></b>		<b>O<sub>4</sub></b>

In this sample table, respondents are assigned randomly (R) to Group 1 or Group 2 at Time 1; at Time 2 (sometimes called “baseline”), each group receives a pretest (O<sub>1</sub> and O<sub>3</sub>); at Time 3, Group 1 receives the experimental intervention (X); at Time 4 (sometimes referred to as “followup”), both groups receive the posttest (O<sub>2</sub> and O<sub>4</sub>). In addition, readers will note that Time 3 is blank for Group 2; this does not indicate that Group 2 does nothing at this time. Rather, this cell refers to the activities of the control group during the time when the experimental group experiences “X.” In many cases, the control group experiences the “status quo” or some neutral condition. In other words, Group 1’s experience with the experimental condition will be compared to Group 2’s experience of whatever condition was already in place at the time of the experiment. For example, in a traditional study of the effects of a new substance abuse prevention program, the experimental group would receive the new curriculum and the control group would receive the school’s existing substance abuse program.

### ***Threats to Validity***

Researchers are constantly asking themselves the question “are the results found in my study attributable to the experimental intervention or some other factor?” These “other factors” actually may produce results that could be mistaken for the effect of the experimental intervention. In this report, we refer to these “other factors” that produce spurious results as threats to validity. In program evaluation research, there are a number of threats to validity that researchers should take into account. For example, consider the study of the effects of a new blood pressure medication mentioned previously. If investigators recruit respondents for the experimental group from a health club and recruit respondents for the control group from a smoking cessation program, any differences between the two groups identified at the end of the study could be attributable to characteristics of these individual groups rather than the effect of the experimental intervention—this is called a cohort effect. Although this is an extreme example, the basic concept applies in all research: if the research requires the use of a control group, evaluators must take great care to make sure that the control group is not significantly different from the experimental group on the range of relevant variables.

Research design experts (e.g., Campbell & Stanley, 1963) recognize that threats to validity are quite common in evaluation research. The following list describes 7 specific threats to validity. We will describe these common problems (adapted from Campbell & Stanley, 1963) here so that readers can become familiar with these ideas before we discuss the characteristics of specific research designs that attempt to avoid these knotty difficulties:

1. ***History.*** History refers to the events other than those planned in the experiment that occur during the course of the study and may have an effect on the study’s outcome, thereby threatening the validity of the research. The classic example is the 1940 study<sup>1</sup> in which students were supposed to read Nazi propaganda material and then be tested by experimenters to determine how this experience affected their attitudes. During the course of this study, France surrendered to the Nazis. Thus, the experiment was ruined, since news of this event most likely had a very powerful effect on attitudes but was not controlled in the study; in other words, the experimenters could not tell whether the changes in attitude were the result of their experimental intervention or the news of France surrendering. History can take many forms in an experiment; it can be something as far-reaching as a national sports figure dying from a drug overdose to something as mundane as a new television advertising campaign about pain relievers. Although every aspect of life cannot be controlled in an experiment, there are study designs that take history into account.
2. ***Maturation.*** Maturation threats to validity refer to changes that come about as a result of the passage of time rather than the experimental intervention. For example, consider a program

1 Collier (1944), as cited in Campbell & Stanley (1963).

designed to improve the way fifth-grade students relate to their peers. If students receive a pretest at the beginning of the year, experience the program during the school year, and receive a posttest at the end of the year, any changes that occurred during the year could have been the result of students simply getting older and more socially competent. In addition, maturation refers to changes that might be expected to occur over shorter periods of time. For example, in an experiment that lasts for several hours at a time, respondents will tend to become more tired, more hungry, etc., as the experiment progresses. These factors can compromise the validity of a research project when observations early in the process are compared with observations later in the process.

3. **Testing.** Testing threats to validity refer to the effects that taking a test has on the results of a second or subsequent administration of the same test. In other words, testing occurs when the process of measuring actually has an effect on the thing being measured. For example, students generally improve their scores on the second administration of standardized tests (e.g., the SAT) even when there is no instruction between the two administrations of the test. This improvement occurs because students become familiar with the test format, the amount of time allowed for each section, the question types, etc. We can apply this example to research settings, where investigators often use pretests and posttests. Other forms of testing can occur as well. For example, on the second administration of a personality assessment, respondents are likely to give responses that they think are more “socially acceptable,” once they have had the opportunity to determine what these responses are (by considering their responses on the first test, talking with friends who took the test, etc.). This phenomenon also applies to any test in which respondents can determine what responses will make them appear less deviant or in some way more socially attractive. A final example is the pretest that actually has the effect that the experimental intervention is intended to have. For example, in an experiment on lowering blood pressure, the initial measurement of respondents’ blood pressure may motivate them to change their day-to-day behavior and work towards lowering their blood pressure—independent of the experimental intervention designed to lower blood pressure.
4. **Instrumentation.** Instrumentation refers to changes in the measurement instrument (e.g., the survey) or in the way in which program evaluators collect data that might account for a difference between baseline scores and follow-up scores. In evaluation research, instrumentation usually occurs in relation to data collectors; members of the research team who collect data (e.g., observe classes and assess or categorize students’ behavior) may be more skillful at their jobs the second time around as a result of their initial experiences at the baseline observation. In this sense, instrumentation can be thought of as a *Testing* effect for the researchers. Alternatively, the data collectors may become more tired or more blase as the experiment progresses. Unless these factors are controlled in the research design, they may cause differences between baseline and follow-up scores; program evaluators could incorrectly attribute these differences to changes experienced by the respondents.
5. **Selection.** Selection threats to validity refer to problems that result when researchers select respondents for the control group(s) using a procedure that is different from the procedure they

use to select subjects for the experimental group(s). For example, if a psychology professor requires students in one class to participate in an experiment but creates the control group from volunteers, she should expect these two groups to differ regardless of what program the experimental group experiences. In other words, if the experimental and control groups are not equivalent to begin with, it is not possible to interpret accurately the differences between these groups at the end of the experiment. The best way to avoid this problem is to select all respondents from the same population and randomly assign them to the experimental or control group(s).

6. **Mortality.** Mortality threatens the validity of evaluation research when different drop-out rates for the experimental group and the control group result in differences between the groups that might incorrectly be attributed to the experimental intervention. One example is research on a standardized test preparation course. If respondents are assigned randomly to experimental and control groups, a rigorous test preparation program may cause the least academically successful students to drop out of the experimental group, while there would be no similar stimulus in the control group. Thus, the experimental group would be composed of the brightest students, and would be likely to score higher than the control group on follow-up measures; in this case, this difference would be attributable to the different drop-out rates rather than the test preparation program. Mortality can be a problem even when there is no control group. For example, intelligence tests administered the first year of college and the last year of college might show an improvement in scores over this time period, implying that four years of college increases intelligence; however, this difference may be the result of the least intelligent students dropping out over the course of four years.
7. **Statistical Regression.** Statistical regression, also known as regression to the mean, is a problem that occurs when groups are selected for their extreme scores on a pretest. On most tests, researchers will notice that the respondents with the worst scores on the pretest generally did better on the posttest, and the respondents with the best scores on the pretest generally did worse on the posttest. Although this finding seems relevant, and could be interpreted to be a result of the experimental intervention, this phenomenon is actually the result of a simple statistical principle. The explanation for this phenomenon is as follows: scores on two identical tests taken by the same respondents generally are not perfectly correlated. In other words, respondents do not get the exact same scores on the two tests, but rather each respondent will get a somewhat lower or somewhat higher score the second time, even if the group mean remains the same for the two tests. The reason for these deviations is that there is some degree of error in the test's ability to measure what it is intended to measure. The more extreme the individual score, the more measurement error it probably represents. In other words, some of the low-scoring respondents got low scores by chance, and will "regress to the mean" (get better scores) on the second administration—not because they actually improved, but rather because of the unlikely statistical probability that a random, extreme deviation will be duplicated in the same respondent on the second administration. Thus, if respondents are selected to receive a special program on the basis

of their extreme scores (high or low), they are likely to get scores closer to the overall group mean the second time, even if no instruction occurs between the tests. For example, if school administrators select students who scored lowest on a reading comprehension test for an experimental remedial reading program, these students are likely to score higher on the posttest, independent of this program. The well-known “sophomore slump”—the tendency for exceptional first efforts to be followed by less impressive results—is, in most cases, nothing more than regression to the mean.

### ***Threats to Generalizability***

The threats to validity described above relate to a variety of questions such as “Do these results mean what they appear to mean? Did the experimental intervention have an effect in this study sample, or does some other factor account for the results?” In addition to these very important questions, researchers also must ask themselves how their study’s results may apply to the world at large. In other words, are the results derived from a specific program evaluation project representative of, and thus generalizable to, other settings? Research that is not generalizable to other settings is usually of little value to those outside of the original program setting. Consequently, researchers make careful efforts to ensure that their results can be applied beyond their specific setting, sample, instrument, time, etc. The following list includes three major factors that can act as threats to the generalizability of research findings:

1. ***Interaction Effects of Testing.*** In some cases, the interaction between the pretest and the experimental intervention is the factor responsible for the study’s outcome. In other words, the pretest makes experimental group respondents more (or less) responsive to the experimental intervention than they would be if they received only the experimental intervention. In this case, it is inappropriate to generalize the results to groups that will not receive the pretest. For example, suppose the CEO of a company wants to test the effects of a short film designed to increase employees’ sensitivity to handicapped people. She decides to give one group a pretest, show them the film, and then give them a posttest; knowing the importance of control groups, she creates a control group that will receive only the pretest and the post-test. As she hypothesized, the experimental group shows an increase in sensitivity. However, this outcome does not necessarily indicate that the film *by* itself would have the same effect. In other words, this outcome may be specific to groups that had *both* the pretest *and* the film. If the effectiveness of the film is dependent on being “warmed up” to the topic by experiencing the pretest, simply showing the film to employees in the future may not be effective .
2. ***Interaction Effects of Selection.*** In some program evaluation scenarios, the results of an experiment may be representative only of the specific sample that was studied. In other words, there may be some interaction between the specific sample that was studied and the experimental intervention that would not exist for other samples. This problem occurs most often when a study uses volunteers or, more generally, when it is difficult to get respondents. If it is difficult to get subjects to participate in a study, the researchers will experience many refusals during respondent

recruitment; the respondents who finally agree to participate will probably be more interested in the study, more inclined to make improvements in their lives, etc. These characteristics also apply to volunteers. If investigators conduct experiments with respondents who have special attributes (e.g., volunteers), the results of the research may be applicable only to this narrow segment of the general population.

3. ***Reactive Effects of the Experiment.*** In many cases, respondents' knowledge that they are participating in an experiment will affect the way they respond to tests and to the experimental intervention: they may try to figure out the experimenter's intent, act out, be on their best behavior, or perform in some other way that is not representative of their usual behavior.

If this is the case, then the results of this experiment will represent respondents in those specific experimental conditions but not necessarily the same (or other) respondents in their everyday settings. In some cases, it is not possible to prevent respondents from knowing that they are part of an experiment. However, in educational settings, it is sometimes possible to disguise an experiment by presenting tests and experimental interventions in place of normal exams and curricula, with no special introductions or explanations. In these cases, it is preferable to conduct the experiment without the students' knowledge.

## **Alternative Study Designs: How to Avoid Research Problems**

The discussion above focused on threats to *validity* ("what do these results really mean?") and *generalizability* ("how applicable are these results to other groups, settings, etc.?"). In this section, we will discuss some specific research designs with reference to these problems. Which of these—or other—designs you choose for your research will depend on the type of setting in which you intend to do evaluation research (e.g., school, community), what type of program you tend to evaluate (e.g., curriculum, media campaign), what resources are available, and a variety of other less prominent but still important practical matters. In this section, we will use the research terminology presented before and will make references to the seven common threats to validity and the three primary threats to generalizability described earlier. We will begin with the most simple of research designs and progress to more complicated designs; the first three designs are "pre-experimental" designs, while the last three are true experimental designs.<sup>2</sup>

**The “Pre-Experimental” Design  
Design #1: The “One-Shot” Case Study**

Design #1				
	Time 1	Time 2	Time 3	Time 4
Group 1	X	O	--	--

**Characteristics:** In this design, a single group is exposed to some experimental intervention and then observed.

**Problems:** This design lacks the fundamental requirement of all research - a comparison of some kind. Without at least one comparison of some kind (e.g., pretest to posttest, experimental group to control group), the results obtained at Time 2 of this design are uninterpretable — we have no idea what Group 1 would have been like had it not experienced “X.” This is such a significant problem that the threats to validity presented above are minor in comparison. We present this design as a reference point for the designs that follow and as an example of what to avoid.

**Design #2: The One-Group Pretest/Posttest Design**

Design #2				
	Time 1	Time 2	Time 3	Time 4
Group 1	O <sub>1</sub>	X	O <sub>2</sub>	--

**Characteristics:** This design is equivalent to design #1 with a pretest added.

**Problems:** While this design is better than design #1 because it includes a comparison observation, it is only worth using if resources are extremely limited and nothing better can be done. The potential threats to validity for this design include *History*, *Maturation*, *Testing*, and *Instrumentation*. *Regression* would be a problem if the respondents are selected on the basis of their extreme scores on the pretest. The potential threats to generalizability include *Interaction Effects of Testing* and *Interaction Effects of Selection*. *Reactive Effects of the Experiment* also could be a problem.

<sup>2</sup> We have adapted these classic research designs from Campbell & Stanley (1963). Interested readers are encouraged to review Campbell & Stanley for a more comprehensive and technical examination of research designs.

### Design #3: The Static Group Comparison

Design #3				
	Time 1	Time 2	Time 3	Time 4
Group 1	X	O <sub>1</sub>	--	--
Group 2		O <sub>2</sub>	--	--

**Characteristics:** Two groups are used; the experimental group receives the experimental intervention and a posttest, the control group receives only the posttest.

**Problems:** The main problem with this design is *Selection*; we have no way of knowing whether these two groups were equivalent before the introduction of X. *Mortality* also could be a problem with this design. *Interaction Effects of Selection* could be a potential threat to generalizability.

### The Experimental Designs

#### Design #4: The Pretest/Posttest Control Group Design

Design #4				
	Time 1	Time 2	Time 3	Time 4
Group 1	R	O <sub>1</sub>	X	O <sub>2</sub>
Group 2	R	O <sub>3</sub>		O <sub>4</sub>

**Characteristics:** This design uses an experimental group and a control group, both of which are assigned randomly, receive a pretest, and receive a posttest. This design, the first of the true experimental designs in our list, effectively controls for all seven of the threats to validity.

**Problems:** The main problem with this design is that it does not control for the *Interaction Effects of Testing*; there is no way of knowing whether X would have the same effect if presented without the pretest. In addition, while this design controls for the threats to validity (e.g., *Testing*, *Maturation*, *History*, *Regression*), it does not measure the specific effects of these factors.

For example, the effect of *Testing* is controlled – both groups experience the pretest, and thus the effect of the pretest cannot explain any differences between the groups at Time 4. However, although this factor is controlled, it is not possible to know whether the pretest does have an effect and, if so, what the magnitude of this effect is. In addition, *Interaction Effects of Selection* and *Reactive Effects of the Experiment* may be problematic; whether they are or not depends on the nature of the specific research project. We can only suggest that researchers take these factors into account and try to minimize the effects of these influences in their research projects.

### Design #5: The Solomon Four-Group Design

	Design #5			
	Time 1	Time 2	Time 3	Time 4
Group 1	R	O <sub>1</sub>	X	O <sub>2</sub>
Group 2	R	O <sub>3</sub>		O <sub>4</sub>
Group 3	R		X	O <sub>5</sub>
Group 4	R			O <sub>6</sub>

**Characteristics:** This design uses four groups to examine every combination of pretest and **X**: (1) pretest with **X**; (2) pretest and no **X**; (3) no pretest and **X**; and (4) no pretest and no **X**. Like Design #4 above, this design control for all of the threats to validity; however, this design offers several additional benefits not available in Design #4. Primarily, this design identifies both the main unique effects of *Testing* as well as the *Interaction Effects of Testing*. In addition, this design identifies the combined effects of *History* and *Maturation* (it is not possible to identify the unique effects of these two factors). Furthermore, the effect of **X** can be identified in four different ways.

**Problems:** This design has no problems per se; the main drawback associated with this design is that it requires more resources than any other design presented in this guide. In addition, as described above, the *Interaction Effects of Selection* and *Reactive Effects of the Experiment* remain potential problems, as with any experimental design.

### Design #6: The Posttest-Only Control Group Design

Design #6				
	Time 1	Time 2	Time 3	Time 4
Group 1	R	X	O <sub>1</sub>	--
Group 2	R		O <sub>2</sub>	--

**Characteristics:** This design is equivalent to Design #3 with the addition of random assignment of respondents to groups. Like the two other true experimental designs presented here, this design controls for all of the threats to validity (the random assignment procedure controls for the *Selection* problems that are characteristic of Design #3). In addition, since there is no pretest, the *Interaction Effects of Testing* are not an issue as they are in Design #4.

**Problems:** Like the Solomon Four-Group strategy, this design also has no problems per se. However, its main drawback is that it is not as comprehensive as the Solomon Four-Group design (i.e., Design #5). For example, this design does not identify the main effects of *History* and *Maturation*, the main effects of *Testing*, or the *Interaction Effects of Testing* as Design #5 does (although it controls for these factors). In cases where researchers suspect that the pretest in itself may be a powerful stimulus to change, they may want to identify the unique effect of the pretest using Design #5. Although Design #5 is undoubtedly the most comprehensive and rigorous design, Design #6 delivers “the most bang for the buck” in most circumstances.

## Preparing to Conduct Research: Steps to Planning and Implementing Evaluation Projects

Research projects involve many steps, and sub-steps. Hulley and Cummings (1988) elegantly summarized the activities of a successful research project in the following six steps:

1. Choose a research question or questions;
2. Develop a research protocol;
3. Pretest and revise the protocol;
4. Conduct the study;

5. Analyze the data;
6. Draw and disseminate the conclusions;

Readers should note that only one of these six steps—step #4—involves actually conducting the study, and that three of these six activities—steps 1 through 3—deal with preparation for conducting the study. Inexperienced researchers tend to put most of their efforts into the development of the program to be evaluated; they tend to think that a good program of prevention or intervention will be sufficient to yield a good evaluation project, and that the details of the research will sort themselves out. This is not the case. Researchers who are likely to overlook a number of important factors often find out later that their evaluation results are useless, and that all the time they have invested in the evaluation project has been wasted. All good research studies have one aspect in common: a great deal of thought and effort went into the planning and preparation of the project. If a study is designed well, it will be much easier to execute. In addition, the results will be much more likely to be meaningful, both to the researcher and to the community at large. The more work researchers put in during the planning stages—identifying the purpose of the project, identifying the research questions to be asked, etc.—the easier the rest of the project will be to complete and, more importantly, the more valuable the research findings will be for all interested parties. Ideally, all of the steps in the research protocol (e.g., selecting respondents, pre-testing) should be developed in detail before the project begins.

Once researchers carefully complete the six steps above, new questions often emerge for consideration, and the research cycle continues again. In the following sections, we will examine each of these primary steps in the research process.

### ***Determining the Research Questions***

As we discussed previously, the first step of a program evaluation is to identify explicitly and precisely *what exactly is the goal of the prevention, education, or treatment program*. Once the purpose of the program has been identified, and its objectives have been specified, it is possible to formulate hypotheses about the outcome of the evaluation research. It is important that these research hypotheses match the purpose of the program. For example, suppose the purpose of a program is to decrease substance use behavior among middle school students. If this is the case, it is irrelevant to offer as a primary hypothesis that the program will change students' attitudes toward substance abuse; the relationship between attitude and substance use is secondary to the goals of the program. It is not worth the considerable effort involved in conducting a well-designed evaluation when the primary hypothesis is secondary to the program objectives while the primary program objectives are ignored. In the example cited above, prevention research would be based on a very common but faulty assumption – that attitude changes will translate to behavior changes. Research already conducted on this issue reveals that positive changes in attitudes and knowledge about substance use are not sufficient to cause a corresponding reduction in substance-using behaviors (Botvin & Botvin, 1992).

### ***Developing the Research Protocol***

The planning stage of the research project involves the development of a research protocol. The protocol development phase of evaluation projects is very demanding and involves careful planning; this protocol will include a thorough and detailed description of each aspect of the research project. The development of the research protocol is a vital part of a research project, since it will guide each step of the implementation of the project. In this section, we will describe and discuss each aspect of the research protocol. These aspects include the following: (1) building an evaluation instrument; (2) identifying the study population; (3) identifying the appropriate research design; (4) power analysis; (5) determining the sampling methods; (6) data coding and entry; (7) data analysis; (8) interpreting the data.

#### **Building an Evaluation Instrument**

Once you have identified the purpose of the program explicitly and developed the corresponding hypotheses necessary to guide the development and implementation of the evaluation project, you should begin working on the instruments that will be used in the evaluation research (e.g., surveys, pretests and posttest). The development of the instruments should follow directly from the specific hypotheses that you will investigate during the evaluation. In other words, the task is to identify the specific survey questions, items, or scales that will measure the changes specified in your hypotheses. It is usually helpful to conduct a literature review to identify other studies that have investigated the same topic—often the authors of these existing studies already will have developed survey items or entire instruments that you can modify for your specific purposes. In general, it is preferable to use survey items that have quantitative response scales. Quantitative response scales are more useful than open-ended items that will require additional interpretive ratings later. For example, consider the following two items:

**(A)** During the past 7 days (including today), how much were you distressed by nervousness or shakiness inside? (0) not at all; (1) a little bit; (2) moderately; (3) quite a bit; (4) extremely<sup>3</sup>

**(B)** How nervous have you been during the past week?

Item A is preferable to item B. Item A above has a quantitative response scale: the response options numerically represent the degree of distress respondents might feel, from 0 (no distress) to 4 (extreme distress). Respondents select the specified point along the continuum that best represents their experience. This type of scale makes comparisons easy. For example, researchers can compare respondents' scores on this item at one point in time to respondents' scores on the item at another point in time. Item B is a qualitative (or "open-ended") item, and is less useful for evaluative research purposes. Although it is possible to compare respondents' written descriptions of their nervousness at one point in time to another description from a different point in time, this task becomes much more difficult with large numbers of respondents and multiple observations. In addition, statistical analysis requires data that is in numerical format. Thus, at the minimum, the researcher using open-ended items would have to read all the responses, create relevant categories,

This item was adapted from the Symptom Checklist-90-R (Derogatis, 1993).

and indicate which categories represent the various responses. Clearly, the use of qualitative items from the outset will be more efficient and provide more reliable data.

### *Survey development hazards*

There are a number of common hazards that researchers often overlook in the process of survey development. The following list presents and explains some of these common mistakes:

- **Items do not specify a time frame.** For example, “How much money do you spend on your favorite hobby?” This item does not specify whether the researcher is referring to the past month, the average month during the past year, the respondent’s lifetime, or some other time frame. Thus, each respondent will decide what time frame they want to use, will not specify a time frame at all, or will be confused and will skip the item.
- **Two questions in one.** For example, “Have you ever gambled in a casino or on the lottery?” In almost every case, the researcher will obtain better data by addressing only one issue in a question. When two items are combined into a question by means of an “and” or an “or,” it becomes impossible to determine to which of the activities the subject was responding.
- **Overlapping response options.** For example, “How long have you worked in your current job? (1) 0-1 year; (2) 1-2 years; (3) 2-3 years; (4) 3 or more years.” In this case, respondents who have worked for one year, for example, will not know whether to select option 1 or option 2; the variation that results will harm the validity of the data.
- **Unbalanced response scales.** For example, “Please indicate how much you agree with the following statement using this response scale: (1) disagree completely; (2) disagree somewhat; (3) neutral; (4) agree completely.” In this example, respondents are given the option of disagreeing somewhat (#2), but not the option of agreeing somewhat. Thus, anyone who would have chosen “somewhat agree” if it had been available will instead choose either “neutral” or “completely agree,” which will skew the results. Ideally, a response scale of this type will have a middle position (usually a “neutral” or “moderate” response) and an equal number of responses in the “positive” and “negative” directions.
- **Other incomplete response scales.** Researchers must do everything possible to ensure that they obtain valid and complete data. Part of this effort includes making sure that each respondent completes the entire survey. If respondents do not find the response option that applies to them, they will probably skip the item altogether. In this case, the researcher does not know whether the respondent skipped the item for some reason (e.g., refused to answer, didn’t see the item) or if the item simply did not apply to that respondent. Thus, the researcher

does not know whether to treat this item as missing data or as not applicable. For example, with the item “Have you ever received a vaccination for Hepatitis B? (1) yes; (2) no,” a respondent who does not know will probably skip the item. This question would be improved by the addition of an “I don’t know” option. Similarly, researchers include a “not applicable” response option when they think that the question may not apply to certain respondents. Finally, researchers should include an “other” response option whenever applicable. For example, “What type of alcoholic beverage do you drink most frequently? (a) beer; (b) wine; (c) mixed drinks; (d) straight liquor; (e) other; (f) do not drink alcohol.” The goal of a well-constructed data collection instrument is to have each respondent provide a meaningful response to every item, even if the respondent does not have the information the researchers is requesting or if the item does not apply to the respondent.

- **Vague questions.** In general, researchers should make questions as specific as possible. For example, the question “Have you seen *Batman*?” could have a number of interpretations. Is the researcher referring to a) the 1989 movie; b) the 1966 movie; c) the animated television series; d) the action figure; or some other version? Researchers should avoid ambiguity and confusion wherever possible by making each question as specific as possible.

### **Identifying the study population**

In most research, it is not possible for investigators to study every member of the population to which they want to generalize their research. As a result, researchers usually select a sample of respondents from the larger population of potential respondents. For example, if a state substance abuse treatment system wants to know how many people will require its services during a one-year period, it does not survey every state resident to obtain this information. Instead, it selects a random sample of the state population and then uses the results to estimate the state total. If sampling is done properly, the sample (i.e., the people who participate in the study) will be representative of the population (in this example, the entire state population). Another example is a research team that wants to develop estimates of the prevalence of substance use among middle and high school students in the United States. In this example, the population being studied is all middle and high school students in the United States. To obtain a representative sample, the researchers will have to consider a range of variables, such as region of the country, type of school (public or private), town size, socio-economic status, etc.

These examples are provided to illustrate the point that the identification of the population being studied is the first step in the process of obtaining a study sample. Once investigators have clearly identified the population to which they want to generalize their results, they will then be able to address the variables of their sampling procedure that relate to generalizability.

## Identifying the Appropriate Research Design

We discussed the issues involved with identifying the appropriate research design in detail in the previous section titled *Alternative Study Designs: How to Avoid Research Problems*. Nevertheless, the task of selecting a specific design can be daunting. When selecting which research design to use for a particular research project, investigators should consider the population to which they wish to generalize their results, the nature of the prevention or intervention program being studied, the threats to validity and generalizability that might apply to their specific research project, and the resources available to them.

### Power Analysis

All research protocols should include a statistical power analysis.<sup>4</sup> The basic purpose of a power analysis is to determine the number of respondents necessary to identify any meaningful differences that exist among the groups being studied. These meaningful differences are referred to as the “effect” of the experimental intervention. Power represents the probability (i.e., percentage likelihood) that a particular research design will identify any existing effect, given a particular number of subjects, and a particular level of statistical significance. For example, a particular research design, with a particular number of respondents, for a specific program effect, at a particular level of statistical significance, can be said to have “80% power.”

Program evaluators can think of a power analysis as a type of research warranty. That is, by determining the power of a particular research design, researchers are making an estimate of the likelihood that they will find an effect, if it exists. Small program or treatment effects are more difficult to identify than larger effects. Therefore, compared with moderate and large effects, small effects will require larger sample sizes to maintain the same level of power, given a constant statistical significance level (i.e., threshold at which researchers will determine that the study groups are meaningfully different). Determining this decision-making threshold, or significance level, is a decision researchers must make based upon many different factors. For example, investigators evaluating a new low-cost, low-risk treatment intervention might be willing to accept a difference between study groups at a lower statistical threshold than they would use in a study comparing well-developed, expensive, or very risky treatments. In the first instance, the significance level will be set more liberally than in the latter (e.g., .05 and .01, respectively).

The calculation of study design power involves five primary factors: (1) the expected size of the program effect (e.g., small, medium, or large); (2) the number of subjects per study group; (3) the statistical significance level that will serve as the threshold for decision making; (4) the research design (i.e., the number of groups that will be compared); and (5) the specific type of statistical test that will be used to analyze the data. Typically, investigators will determine all of the

<sup>4</sup> For complete descriptions of power analyses, see Cohen (1988), Fleiss (1981; 1986), and Kish (1965).

factors above except the number of subjects per study group; they will then use this information to determine how many respondents per group are needed to achieve a desired level of power (e.g., 80% is a frequently used convention). While a high level of power is generally desirable, too much power can present problems for evaluators. Excessive power can identify very small, and perhaps not very meaningful, differences between the experimental and control groups. If an expensive and difficult-to-implement treatment or prevention program hangs in the balance pending the outcome of an evaluation project, investigators should usually avoid excessive power. Although a comprehensive review of power analysis is beyond the scope of this paper, readers should refer to Cohen (1988) and Bornstein, Rothstein, & Cohen (1997) for more detailed explanations of these issues.

## Determining the Sampling Methods

Once investigators have conducted a power analysis to determine the necessary sample size, they must determine what sampling methods they will use to obtain respondents for their study. On the most basic level, sampling methods are divided into two categories: probability (i.e., random) sampling and nonprobability (i.e., nonrandom) sampling. Each of these categories includes several specific sampling designs. For example, probability sampling includes four specific sampling designs: (1) simple random sampling; (2) systematic sampling; (3) stratified random sampling; and (4) cluster sampling. Nonprobability sampling includes three specific sampling designs: (1) consecutive; (2) convenience; and (3) judgmental. We will briefly summarize each of these approaches to selecting study samples (adapted from Hulley, Gove, Browner, & Cummings, 1988).

### *Probability Sampling*

(1) ***Simple random sampling:*** With simple random sampling, the investigator first identifies and enumerates every member of the population to be studied; the investigator then randomly selects a certain number of respondents from this list. This design, while very methodologically sound, can be difficult to execute if it is troublesome to identify and enumerate all of the members of the population of interest.

(2) ***Systematic sampling:*** With systematic sampling, as is also the case with simple random sampling, the investigator first identifies and enumerates every member of the population to be studied. However, rather than selecting respondents randomly, the investigator selects a random starting place on the list and then selects every *n*th person on the list (e.g., every other person, every third person). This sampling design leads to problems when there are patterns in the order in which members of the population are listed. In general, if investigators are able to accomplish the first step of this procedure (i.e., identifying and enumerating every member of the population), it is preferable to use simple random sampling.

(3) ***Stratified random sampling:*** With stratified random sampling, the investigator first identifies and enumerates every member of the population; the investigator then divides the population into subgroups (i.e., strata) and randomly selects a portion of the study sample from each of these strata. In most cases, these strata are based on gender, race, or some respondent characteristic that is either underrepresented in the population or essential to represent equally in the study sample (e.g., age groups). Randomly sampling from stratified groups allows the investigator to analyze data from underrepresented subgroups of the population with appropriate power.

(4) ***Cluster sampling:*** Cluster sampling is useful when it is not possible or feasible to identify and enumerate every member of the population. With cluster sampling, the investigator instead identifies and enumerates existing groups, or clusters, of respondents (e.g., schools, homes, treatment programs). The investigator then randomly selects a certain number of these clusters and includes all members of the selected clusters in the study. A variation on this procedure is two-stage cluster sampling, in which clusters are selected randomly and then respondents are selected randomly from within the selected clusters. This strategy might be appropriate in, for example, a study of middle school students within a particular state. In this instance, rather than identifying and enumerating every middle school student in the state, the investigator would identify every middle school, randomly select a specific number of these schools, and then randomly select students from within the selected schools. The potential hazards of this design relate to the tendency for naturally occurring clusters to be homogeneous; for example, students within one school might all be similar with respect to a particular variable (e.g., they might have a common socioeconomic status), while students in another school would be similar to one another but different from the first school. Thus, it is important to identify clusters that represent a range of values of the variables of interest.

### *Nonprobability Sampling*

(1) ***Consecutive sampling:*** The first step of consecutive sampling is to select a site (or multiple sites) from which respondents can be sampled (e.g., treatment programs, hospitals). These sites should be representative of the population being studied; ideally, the investigator would select these sites randomly. The investigator then includes in the sample every available member of these sites over a particular time period. This method is most often used in treatment settings, where investigators include in the study each patient admitted to the selected treatment site(s) over a particular period of time. Although this method is often more practical than the probability sampling methods described earlier, and is the strongest of the nonprobability designs, there are a number of hazards associated with this design. The primary hazard associated with this design is that investigators often select sites for their convenience rather than for how well these sites represent the population of interest (see Convenience sampling below). Another potential hazard with this design is that the period of time during which respondents are admitted to the study may be too short and therefore may not accurately reflect longer-term patterns.

For example, characteristics of respondents who enter an inpatient substance abuse treatment facility during the winter may be different from the characteristics of respondents who enter in the summer. If the treatment facility bases a study on all December and January admissions, the results of this study may not be representative of the year-long pattern.

(2) **Convenience sampling:** Convenience sampling involves the selection of respondents who are readily available or feasible to include in the study. Although the benefits of this design in terms of cost and logistics are obvious, we encourage evaluators to avoid this design. In most cases, the potential benefits are outweighed by the lack of representativeness of the derived sample. As we discussed above in the *Interaction Effects of Selection* section, volunteers and other readily accessible members of the population often have specific characteristics that make them unrepresentative of the population the investigator wishes to study.

(3) **Judgmental sampling:** In most cases, judgmental sampling is a variation of convenience sampling. With judgmental sampling, investigators select respondents who they consider to be appropriate candidates for the experimental intervention. For example, managers of a substance abuse treatment facility may identify specific clients in the facility for whom they believe a new treatment modality is appropriate and select these clients for an evaluation of this treatment modality. This design has hazards similar to those of convenience sampling described above.

## **Response Rates**

Every consideration of the sampling methods described in the previous section should include careful consideration of response rates. The response rate of a study is the percentage of the eligible respondents that actually participated in the study (e.g., completed the program, completed a survey). In their research protocol, evaluators should include the specific minimum response rate that they will consider acceptable for their project. Considering 70% as the minimum acceptable response rate is a general convention that investigators can use as a guideline. Although a 70% response rate is generally considered to be the lowest level acceptable, readers should note that even a 75% response rate may be associated with difficulties. For example, “A non-response rate of 25%, although a good achievement in many settings, can seriously distort the observed prevalence of a disease when the disease itself is a cause of non-response” (Hulley et al., 1988, p. 27). What response rate the investigators obtain will determine how many respondents they will have in their study, and, as the earlier section on *Power Analysis* indicates, the number of respondents is an important consideration for data analysis. The potential number of respondents produced by a particular level of response should be anticipated and planned carefully. We cannot overstate the importance of response rate considerations for evaluation research. A low response rate can, in effect, turn what would have been a representative sample (e.g., a random sample) into a nonrepresentative sample of volunteers. In other words, when a low response rate is obtained, it is likely that those who were willing to participate are meaningfully different from those who

refused to participate on a range of characteristics. Readers should see the previous section titled *Interaction Effects of Selection* for more information on this topic. Response rate is an integral part of a research protocol that has implications for power analysis, interaction effects of sampling, and a range of other issues.

## Calculating Response Rates

The proper way to calculate a response rate is as follows:

$$\frac{\text{number of respondents participating in the study}}{\text{total number of respondents eligible to participate in the study}} * 100$$

Often there is confusion about who exactly is an eligible respondent. All respondents whose participation in the study is requested by the investigators are considered eligible respondents. For example, students who were selected as part of a study sample but were absent on the day of survey administration are considered eligible respondents and must be included in the calculation of response rate. Therefore, the research protocol should address the issues related to absentees and to converting non-responders (e.g., refusals) to responders. If evaluators prepare a research protocol that includes estimates of the percentages of respondents they expect to be absent or they expect to refuse to participate, as well as plans for follow-up efforts to recruit with these non-participating but eligible respondents, the investigators will be much more likely to achieve a satisfactory response rate. For a more thorough discussion of the proper calculation of response rates, and the calculation of compound response rates for cluster sampling designs, readers are referred to Frankel (1982).

## Planning Data Coding and Entry

A research protocol should specify precisely how data will be collected, and then how it will be transferred to an analyzable state. In most cases, researchers will enter respondents' data (e.g., survey responses) into a computer program (e.g., SPSS, SAS) to allow for analysis. However, before researchers can enter data, they must "code" it. That is, they must identify each unique variable in their data set and assign numerical codes to each of the response options. This transformation will allow the data to be analyzed statistically. Including information on data collection, coding, and entry in the research protocol (and then following this protocol) is vital to data integrity. During the process of data entry, the quality of data should be monitored carefully. A good guideline to follow is that each time data is transferred to a new format, its integrity should be monitored. For example, data should be checked when respondents complete their surveys, and again when investigators translate the data from the surveys to the computer database. We suggest that a research monitor randomly select a minimum of 10% of the cases in each instance and compare the initial data with the resulting data to identify any errors. The research team should report the error rates that result from these comparisons – ideally, these comparison will yield an error rate of less than 1%. Finally, and of great importance, evaluators must make careful provisions to assure the confidentiality, and, when appropriate, the anonymity, of respondent data.

The research protocol should anticipate how investigators will analyze the obtained data. As described before, this information represents an integral part of a power analysis. This information is also directly related to the investigators' hypotheses and the development of their instrument. Readers should view all these factors as closely interconnected—specific hypotheses will lead to the need to collect specific data (i.e., the items on the instrument), and to analyze this data in specific ways. These interactive factors guide the conduct of a power analysis, which indicates the number of respondents that will be necessary. Finally, considerations about the population of interest and the number of respondents necessary to represent this population adequately will determine the actual study sample that investigators choose to examine.

Investigators should consider the variety of descriptive, parametric, and nonparametric data analysis techniques that are available. When only small samples are available, researchers should consider nonparametric statistics as the analytic tools of choice. When larger samples are available and the variances of the study groups found to be homogeneous, researchers can consider using parametric statistics. Research protocols should not simply list a medley of statistical instruments as the analytic choices. Data sets should be diagnosed and evaluated for the proper statistical “fit.” Investigators should provide a rationale for using specific statistical devices and be certain that their obtained data set matches the requirements of these statistics. In other words, when characteristics of the data set fail to satisfy the specific assumptions of a statistical test, other statistics should be applied. We strongly encourage evaluators who are not familiar with data analysis and statistical techniques to consult a statistician during the protocol development stages of their research projects—long before data is collected. This planning procedure will help assure that the statistician can complete the necessary data analysis when the time comes.

### **Planning for Data Interpretation**

To ensure that a survey protocol includes the sampling and design elements necessary to meet each investigator's unique data needs, we suggest that the proposed survey protocol consider how the study results will be applied once the analyzed data is obtained. What action will the investigators take if their hypotheses are supported? Alternatively, what action will investigators take if their hypotheses are not supported? What are the critical planning, policy, or funding decisions that face the investigator? Who will be involved in the development of new programs? Which jurisdictions will use the survey results for planning purposes? Each of these and other key questions will assist investigators in making sampling decisions that will assure adequate respondent samples in key regional or respondent attribute strata.

### ***Pilot Testing and Revising the Protocol***

Investigators will find that Murphy's Law—*whenever something can go wrong it will*—is especially true in research settings. This is the case because of the necessary complexity of any

research project. As a result, it is always advisable to pilot test the research protocol. Investigators can conduct a pilot test by following their data collection and entry procedure on a small number of subjects. In most cases, 15 to 20 pilot respondents is an appropriate number. Investigators should keep in mind that these pilot respondents should be representative of the population from which the actual study sample will be drawn – for example, if the study sample will be *n*<sup>th</sup> grade students, investigators should not pilot test the protocol on adults. In addition, pilot respondents should not be eligible for participation in the actual study (cf. the problems caused by *Testing* in the section on *Threats to Validity* above). This pilot test procedure will identify problems with instructions to respondents, problems with the survey instrument, and other improvements that investigators can make in the data collection and entry protocol. If investigators fail to conduct this pilot test, they will most likely find that they have to discard data for entire survey items that turn out to be ambiguous, poorly worded, or for some other reason unintelligible to the respondents.

### ***Implementing the Research Program***

As readers have learned, a properly developed research protocol will leave very few procedures open to guesswork or last-minute planning. If the investigators have developed the research protocol properly, they will find that the implementation of the study runs smoothly. However, there are a number of practical issues involved in conducting a research project that are not addressed in the research protocol. For example, researcher will learn that it takes extra effort to develop and maintain the personal and professional relationships that are necessary to conduct research. Since, in the majority of cases, investigators will be exposed to unfamiliar settings as part of their research (e.g., school systems, communities, treatment programs), they should make every effort to discuss the research goals and the potential for improvement with the administrators of these organizations. The more contact investigators have with the organizations among which they are planning to conduct research, the more these organizations' expectations will be aligned with the investigators' expectations, and the fewer problems investigators will encounter in the actual implementation of the study.

A related practical topic is the training of the research project personnel. The personnel who actually collect data (e.g., administer surveys, conduct interviews) are particularly important to the project, since the quality of the data is dependent on data collectors' following the proper protocols accurately and consistently. The problems associated with *Instrumentation* discussed in the *Threats to Validity* section above are often directly related to the quality and consistency of data collectors. Thus, all data collectors should be trained to recognize and avoid *Instrumentation* problems as well as any lack of consistency or deviation from the established protocol. In many cases involving school-based research, teachers will be responsible for collecting data – in fact, this scenario is generally preferred, since it helps to avoid problems associated with *Reactive Effects of the Experiment* (see *Threats to Generalizability* section above). In this case, training is very important, since many teachers will not have had experience with research projects. These training issues also apply to those who actually implement the experimental intervention (i.e., the program being

evaluated) –in school-based research, teachers most often fill this role as well. Similarly, the training of data entry personnel should include explanations of the importance of their work and standard to which they will be held (e.g., less than 1% error).

### ***Drawing Conclusions***

All of the issues presented in this report relate to identifying the effect of a particular program and the generalizability of these findings to other settings. An additional issue that should be an integral part of any evaluation project is the time frame of the effect – for example, if a program brings about a particular change among the respondents, how long does this change endure? With respect to the time frame of program effects, investigators can only draw conclusions about the time frame they study. In other words, if a posttest conducted one month after the completion of a program shows a change among the respondents, the investigators can conclude that the program results in a short-term (i.e., one-month) change. It would not be appropriate to generalize this finding to longer time frames. Thus, investigators should attempt to conduct their studies over the time frames that relate to the phenomenon of interest. For example, “resistance skills training” programs designed to delay the onset and decrease the prevalence of smoking among adolescents has been shown to be effective in the short term – these programs have been shown to reduce the onset of smoking by 35% and 45% and to reduce the prevalence of smoking by 43% to 47% *after the initial intervention* (Botvin & Botvin, 1992). However, long-term (e.g., 4-5 years) follow-up studies have shown that these changes are not maintained over time, indicating the need to extend the intervention or provide periodic “booster sessions” to reinforce and extend the program effects. If these programs had been studies only over the short-term, investigators may have erroneously concluded that a single intervention was sufficient to change behavior meaningfully.

## **Making Program Decisions: Using Evaluation Research**

One of the most important challenges to program evaluators is the tendency of administrators to disregard scientific evidence. Personal preferences and face validity often drive the implementation and maintenance of community treatment and prevention programs. The real value of any evaluation surfaces when it is used to direct the implementation, maintenance, or modification of the program that it examined. Unfortunately, many well-done program evaluation studies have ended up on a researcher’s shelf collecting dust and influencing no one. Perhaps the primary reason for this state of affairs is that evaluators often think their job is complete when the report is distributed. Nothing could be further from the truth. Evaluators must remain available to help administrators understand and interpret the results of their research. In addition, evaluators must join with administrators to institutionalize evaluation and the programs that were under study.

Once administrators receive the results of a program evaluation, they are faced with matters of interpretation and application. That is, once information is available about the nature of a program's effectiveness, administrators must decide what to do with their program. For example, let's examine the experience of school systems with the DARE (Drug Abuse Resistance Education) program. Not long ago, Ennett, Tobler, Ringwalt and Flewelling (1994) reported that DARE programs were found to influence students' drug knowledge and social skills. However, these changes were less than those obtained from more interactive prevention programs (e.g., peer-to-peer teaching) on measures of drug knowledge, drug attitudes, social skills, and drug use. Furthermore, the results of this research suggested "that DARE's core curriculum effect on drug use relative to whatever drug education (if any) was offered in the control schools is slight and, except for tobacco use, is not statistically significant" (p. 1398). Given this finding – which buttresses other similar research – observers might expect that school administrators would cut back using DARE programs. While this was the case in some cities and towns, on balance, schools have not abandoned the DARE program. Although DARE fails to do what it claims as its primary objective – that is, to reduce drug abuse – it succeeds at improving the relationship between police and young people. In addition, it brings school systems and police departments together in a working relationship. Had DARE identified these objectives as program goals, the program would have been evaluated against a different standard and likely been considered a rousing success.

Given the results of a careful program evaluation, administrators must first determine if the program had any effects on those exposed to the prevention or treatment activities. If the effects were negative or neutral, these effects must be weighed against the positive findings. Absent any positive findings, administrators would be wise to reconsider their relationship with the program. However, negative findings that are outweighed by positive results can be viewed as side effects. As is the case with the use of life-saving medications and other necessary health care interventions, the extent and impact of these side effects always must be measured against the positive impact of the intervention. Botvin, Baker, Dusenbury, Botvin, and Diaz (1995) remind us that the effects of even the most positive school-based prevention programs may be fleeting. However, administrators and researchers have learned that by repeating prevention interventions – like booster inoculations – these programs hold considerable potential for success. Readers interested in more comprehensive review of school based substance abuse prevention programs should review the work of Ellickson (1995).

Next, when a program's positive effects reside in a domain other than those program developers claims, administrators must reconsider the specific goals of the program under examination. Did the program fulfill its stated purpose? If not, did it fulfill another unanticipated but nevertheless important purpose? Absent specific positive effects that administrators can identify, the search for new programs must continue. However, when an evaluation of the consequences of program suggests that its effects are positive, administrators—hopefully encouraged by program evaluators—must begin the very difficult task of stabilizing the activity within the relevant institutions, and building an ongoing evaluation component to monitor these efforts.

## References

- Borenstein, M., Rothstein, H., & Cohen, J. (1997). *SamplePower 1.0*. Chicago: SPSS, Inc.
- Botvin, G. J., Baker, E., Dusenbury, L., Botvin, E. M., & Diaz, T. (1995). Long-term follow-up results of a randomized drug abuse prevention trial in a white middle-class population *Journal of the American Medical Association*, 273(14), 1106-1112.
- Botvin, G.J. & Botvin, EM. (1992). School-based and community-based prevention approaches. In J.H. Lowinson, P. Ruiz, R.B. Millman, & J.G. Langrod (Eds.), *Substance abuse. A comprehensive textbook*. Baltimore: Williams & Wilkins.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally & Company.
- Cohen, J. (1988). *Statistical power analyses for the behavioral sciences*. (second edition). Hillsdale, NJ: Lawrence Erlbaum.
- Derogatis, L.R. (1993). *SCL-90-R. Symptom checklist-90-R*. Minneapolis, MN: National Computer Systems, Inc.
- Ellickson, P. L. (1995). Schools. In R. H. Coombs & D. Ziedonis (Eds.), *Handbook on drug abuse prevention. A comprehensive strategy to prevent the abuse of alcohol and other drugs* (pp. 93-120). Boston: Allyn & Bacon.
- Ennett, S. T., Tobler, N. S., Ringwalt, C. L., & Flewelling, R. L. (1994). How effective is drug abuse resistance education? A meta-analysis of project DARE outcome evaluations. *American Journal of Public Health*, 84(9), 1394-1401.
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. (Second ed.). New York: John Wiley & Sons.
- Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: Wiley.
- Frankel, J.R. (1982). *On the definition of response rates. A special report of the CA SR 0 task force on completion rates*. Port Jefferson, NY: The Council of American Survey Research Organizations.
- Guttentag, M., & Struening, E. L. (Eds.). (1975). *Handbook of evaluation research*. (Vol. 2). Beverly Hills, CA: SAGE Publications.

Hulley, S. B. & Cummings, S. R. (Eds.). (1988). *Designing clinical research. An epidemiologic approach*. Baltimore: Williams & Wilkins.

Hulley, S. B., Gove, S., Browner, W. S., & Cummings, S. R. (1988). Choosing the study subjects: Specification and sampling. In S. B. Hulley & S. R. Cummings (Eds.), *Designing clinical research: An epidemiologic approach* (pp. 18-30). Baltimore: Williams & Wilkins.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons, Inc.

Struening, E. L., & Guttentag, M. (Eds.). (1975). *Handbook of evaluation research*. (Vol. 1). Beverly Hills, CA: SAGE Publications.