EVALUATING TRAFFIC SAFETY PROGRAMS: A Manual for Assessing Program Effectiveness

Fredrick M. Streff, Ph.D.
The University of Michigan
Transportation Research Institute

ACKNOWLEDGMENTS

This manual was prepared with the support and cooperation of the Michigan Office of Highway Safety Planning and the U.S. Department of Transportation, National Highway Traffic Safety Administration.

A special acknowledgment and thanks to Kathy Crockett Richards for her illustrations and the manual's design and layout.

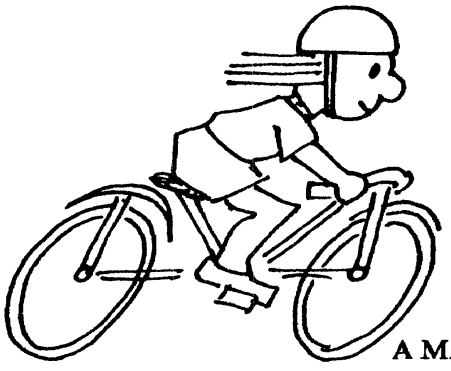
CONTENTS

CHAPTER 1
Why Evaluate?

The local Women's League decided to start a series of educational programs to encourage students in the elementary school district to use bicycle helmets when riding to and from school and at play. A set of three instructional class sessions is designed to be presented to the students during P.E. classes in the early spring. Working with the school district, the classes are developed and held as scheduled in the spring. The teachers report the kids really liked the materials presented.

This is typical of many (but not all) community traffic safety efforts. Much time and effort is spent designing and carrying out programs, but evaluation often relies on anecdotal reports of how people liked the programs. Reports of how programs are received are useful. However, because the effects of the program on student helmet use were not systematically evaluated, the Women's League and the school district cannot determine if the program had its intended effect. Thus, decisions about modifying or repeating the program as originally designed must be made without information about the program's true effectiveness for increasing bicycle helmet use by elementary school children.

Evaluation is a critical but often overlooked component in the development and implementation of traffic safety programs. This is especially true for community traffic safety efforts that are generally run with volunteer staffers and small budgets. However, programs run with limited resources are the types of programs that can benefit most from evaluation. Effective evaluation provides information necessary to help concentrate available resources where they have the most effect.



Empirical evaluation findings also can be used to support programming decisions and funding requests to program sponsors. Sponsors typically want to know the outcome of projects they support, and effectiveness evaluations provide much of the information they require. Program effectiveness results based on a thorough

evaluation are valuable for supporting requests for funding to continue successful programs, as well as modifying or enhancing programs that didn't achieve all they were expected to. The more thorough the evaluation, the more useful it will be for supporting proposed program changes.

Program participants are generally interested in the outcome of the program in which they were involved. Systematic evaluations can help to answer questions they may have, and may encourage their future participation. Another important use of evaluation data is public relations. That is, distribution of program evaluations to the media. These data may be used to promote the success of a project (e.g., *let's congratulate the PTA for increasing safety belt use at the high school by 70%*), to enlist further support for a project (e.g., *while a designated driver program has been set up by over 75% of the bars in the county, we hope more bars will participate when they find out how effective the program is*), or as a component of an on-going project (e.g., *to date, we have increased arrests for speeding and drunk driving by 50% during our special speed and drunk driving enforcement patrols program, and expect to continue to vigorously enforce these laws.*)

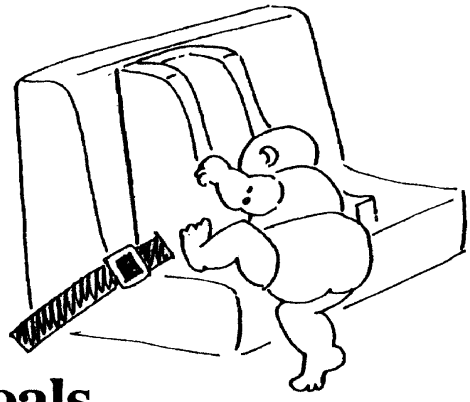
There are many evaluation strategies to provide information to answer questions important to evaluating program effectiveness. Evaluations of program effectiveness should be able to answer such questions as:

- ✔ Was the program conducted as planned?
- ✔ Did the program reach the intended target group?
- ✔ Did the program achieve what it intended to (e.g., change behavior or attitudes, educate people)?
- ✔ What did the program cost (in terms of dollars, person-hours of effort, donated materials)?

Please keep in mind that evaluation is useful for assessing more than program effectiveness alone. Evaluation also should play a role in program development and implementation. Pre-program evaluations can help to determine the nature and magnitude of problems a program may attempt to affect (e.g., the number of alcohol-involved crashes and deaths in your county may be high relative to the state or neighboring counties). Pre-program evaluations also can help determine where and among what groups problems are greatest (e.g., many of those alcohol-involved crashes may be found to occur on rural two-lane highways at night involving vehicles with male drivers under age 25).

While pre-program planning evaluations are important, the focus of this

manual is on the design of practical evaluations for assessing program effectiveness. The manual is divided into five major parts. Identification of program components and goals is described in Chapter 2. This identification process is crucial because evaluation strategies are determined in large part by the nature of the program to be evaluated. Components necessary for designing an effective evaluation plan are discussed in Chapter 3. This includes determination of the purpose for the evaluation, evaluation design, methods to measure program effects, selecting evaluation samples, and data analysis. Chapter 4 describes how to communicate evaluation findings to program sponsors, program participants, and the media. Suggestions for identifying resources for assistance in conducting evaluations are described in Chapter 5.



CHAPTER 2

Identifying Program Goals and Components

The Confederation for Appropriate Child Seat Use decided to distribute pamphlets describing proper installation and use of child safety seats as the central project for the coming fiscal year. These pamphlets would be made available to pediatricians and health clinics where children age 0-4 years are commonly treated. The confederation wanted to evaluate the program, but members were unsure what should be evaluated. Was the central project goal establishing an effective distribution system to the doctors' offices, educating parents about proper child seat use, changing attitudes about child seat use, increasing the proportion of child seats used properly, or some combination of these goals? How would the program be implemented? Would the confederation distribute the pamphlets once at the beginning of the year, at regularly spaced intervals, on request from the doctors, or would they check with the offices periodically to determine if they needed more pamphlets?

Effective evaluation planning begins with a clear understanding of the goals and methods used in the program being evaluated. Detailed information on program goals, design, timing and placement are necessary for the development of the evaluation plan.

PROGRAM GOALS

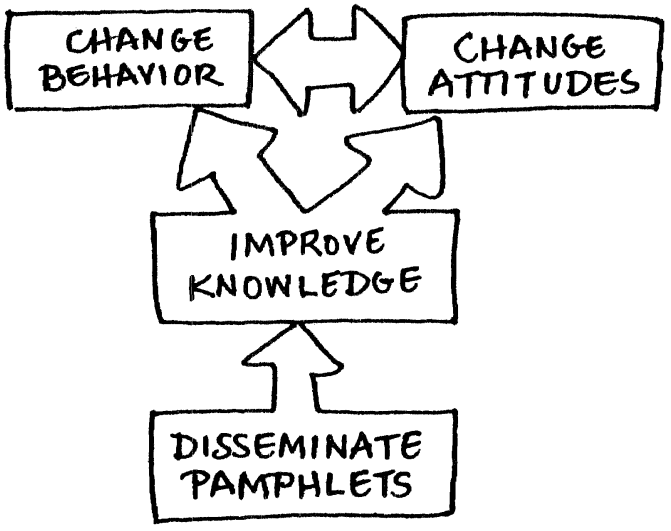
Goals for traffic safety programs usually involve changes in *behavior, knowledge, attitudes or opinions, administration,* or some *outcome* resulting from changes in behavior, knowledge, attitudes, or administration. Usually programs have many goals. Consider the example given at the beginning of this chapter.

The confederation's project involves dissemination of pamphlets describing proper installation and use of child safety seats. Thus, one could say the central goal of the project is *administrative*, that is, developing an adminis-

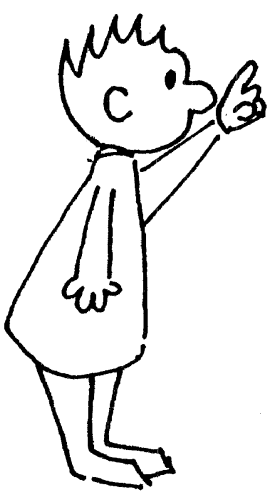
trative system to distribute child seat pamphlets. The purpose of distributing the pamphlets is to educate, or improve parents' *knowledge* of the proper installation and use of safety seats. The goal of improving knowledge is to change parents' *attitudes* toward safety seat use and to change their *behavior* so that seats are installed and used properly.

The ultimate goal of most traffic safety programs is to reduce injuries and deaths resulting from traffic crashes. That is, to change the *outcome* that results from traffic crashes. The most complete evaluation would be able to measure each of these project goals to determine the effectiveness of the project at each step of the way. However, there are always limits to how much and what can be measured and evaluated in a given project. Selecting the program goal or goals to evaluate is a critical step in developing the design for evaluation. This will be discussed in greater detail in the next chapter.

PROGRAM DESIGN



In addition to identifying program goals, you must also identify the design of the program before beginning to develop an evaluation plan. Some programs are one-shot interventions. One-shot interventions are programs where the intervention happens at one, relatively brief point in time. Examples of one-shot interventions include a safety belt promotion booth at a local health fair, a special lecture or movie shown at a school, or a special safety incentive day.



Sometimes programs involve several separate but related programs rather than a single program. Multiple intervention programs are common in communities where an organized traffic safety effort is in place. Multiple intervention programs may focus on a single specific target (e.g., safety belt use) or may focus on a more general target (e.g., alcohol-impaired driving).

A multiple intervention safety belt promotion project may

involve school-based interventions (e.g., a high school safety belt use contest), industry-based programs (e.g., incentive programs offered by local companies for their employees), as well as enhanced police enforcement of safety belt use laws. The goal for each of these programs is to increase safety belt use in different environments.

A multiple intervention program for a general target such as alcohol-impaired driving generally involves several programs designed to change a variety of related behaviors, attitudes, or knowledge in different environments in different ways. Such a multiple intervention program might include promoting use of designated drivers, establishing "tipsy taxi" services to provide rides for people who may have had too much to drink, increasing enforcement of laws prohibiting service of alcohol to minors and intoxicated patrons, and using sobriety checklanes. While the ultimate outcome in these programs is a decrease in alcohol-impaired driving, resulting in fewer crashes and injuries, each individual program in the larger package has a different specific target. Designated driving programs are implemented to ensure that at least one driver in a group remains sober. "Tipsy taxi" services provide rides for impaired patrons when there isn't a sober driver available in a group or the impaired person is alone. Enforcing alcohol sales laws reduce sales to particular problem groups to prevent intoxication. Checklanes act to deter persons who may have had too much to drink from driving for fear of being caught and arrested.

Checklanes act to deter persons who may have had too much to drink from driving for fear of being caught and arrested.

Evaluation of multiple interventions is often conducted at two levels. First, multiple intervention programs can often be considered as several, one-shot interventions, and evaluated as such. *How was belt use affected in each of the intervention sites? How many people took a ride from the "tipsy taxi" service? How often were minors sold alcohol?* These evaluations should be designed separately to take advantage of the unique situations surrounding each individual intervention.

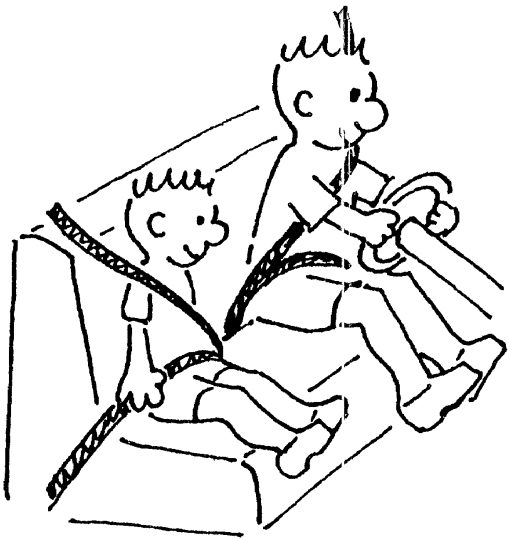
Second, there should be an attempt to determine the effect of the multiple intervention program taken as a whole. *How did the safety belt promotion package affect belt use in the community at large? Did the alcohol-impaired driving package reduce the incidence of crashes involving drivers who had been drinking?*

Often overlooked in evaluation are administrative programs, or administrative components in programs. Rather than overt interventions designed to cause a specific change in attitude, knowledge, or behavior, some programs or program components are administrative in nature. For example, a program may be designed to create and implement a community traffic safety committee, a drunk driving task force, or a similar administrative organization. It is important to evaluate how these programs develop and function. This type of evaluation can help to assess changes that may be needed within the organization or how to implement similar programs more effectively in the future.

PROGRAM TIMING and PLACEMENT

Some community traffic safety programs involve one or more interventions implemented over an extended period of time, others are implemented as one-shot interventions or interventions over a relatively brief amount of time. Timing of program implementation is an important factor to consider when developing your evaluation plan. Premature or delayed measurement can result in underestimation of program effects.

Program placement is also a critical program component to consider for evaluation planning. Evaluation plans should take advantage of opportunities created by program placement. For example, a safety belt promotion program at a work site with a parking lot for employees only is ideal for evaluation. Employee belt use can be observed as employees enter and depart the lot. Because it is an employee lot, we know that the people we observe were exposed to the employee belt use program. To the extent possible, program plans should include evaluation in the beginning rather than as an afterthought to take full advantage of opportunities and to avoid problems that may arise from delayed evaluation planning.



CHAPTER 3
Designing an Evaluation Plan

The county sheriff's department and two local police departments have decided to combine their efforts to conduct an enforcement and PI&E (public information and education) campaign to reduce speeding and increase safety belt use on a stretch of highway in the county. The goals of the program are to increase people's knowledge of the safety benefits of belt use and driving at the proper speed, increase public support for the enforcement campaign, reduce the incidence of speeding, increase safety belt use, and to reduce the number and severity of crashes that occur on the highway. Their program is sponsored in part by funds from the state. As a condition of the funding, the departments are required to conduct an evaluation of the project. In addition, the departments want to evaluate the program to provide information to understand the value of evaluation data to fine tune and expand the program. They now face the question of how to design and conduct a thorough and valid evaluation.

DETERMINING THE PURPOSE FOR EVALUATIONS

The first step to designing an evaluation plan is to determine what the purpose of the evaluation is. As discussed in the example, the evaluation of the enforcement/PI&E project may be conducted to meet requirements of the funding agency, to gather data to determine changes that could be made to enhance the effectiveness of the program, to gather data necessary for supporting an expansion of the program to other counties or local police agencies, or any combination of these goals. Identifying the "why" of the evaluation is the first step in the design process. Why a program is evaluated determines in large part the questions the evaluation should be designed to answer. Some of the questions evaluations may be designed to answer are given on the following page.

QUESTIONS
that Could be Asked

BEHAVIOR QUESTIONS:
- Did the program reduce the number of drivers exceeding the speed limit on the highway?
- Did the program reduce the average speed on the highway?
- Did the program increase safety belt use on the highway?
- Did the program affect speeds on other roads in the county?
- Did the program affect safety belt use on other roads in the county?

KNOWLEDGE QUESTIONS:
- Are safety belts effective in reducing injuries and saving lives?
- Can't safety belts actually cause injuries?
- Won't safety belts trap you in the car and prevent you from escaping from a crash involving fire or water immersion?
- Why is speeding a problem?
- Weren't roads designed to handle higher speeds than the current limit allows?

ATTITUDE/OPINION QUESTIONS:
- Did the public know about the program?
- Did the public support the program?
- Would the public support similar programs in the future?
- Did the program affect the public's opinions of traffic law enforcement?
- Did the program affect the public's opinions of police in general?

OUTCOME QUESTIONS:
- Did the number of crashes occurring the project highway (or crash rate) decrease?
- Did the number of speed-related crashes on the highway (or crash rate) decrease?
- Did the rate of injuries or deaths per crash decrease on the highway?
- Was there any change in the number, rate, or severity of crashes on other roads?

ADMINISTRATIVE EVALUATION QUESTIONS:
- How many officers were assigned to the project?
- How many man-hours were spent on project patrol?
- How many citations (speeding, safety belt, other) were issued?
- How was the program managed (e.g., joint task force)?
- How many press releases on the project were issued?
- Date, time, number of other media contacts (e.g., radio, TV, newspapers)?
- Type and location of other PI&E efforts (e.g., roadside signs, posters, table tents, placemats)?

COST-BENEFIT QUESTIONS:
- What is the cost of increased enforcement personnel?
- What is the cost of administering project personnel?
- What are the costs of preparing and distributing PI&E materials?
- What are the costs (savings) associated with changes in crash frequency and severity?

Once you have determined why an evaluation will be conducted, you must determine the specific questions the evaluation will try to answer. Knowing the evaluation questions will guide what data will need to be collected, as well as how the data will be collected, by whom, where, and when. The program described above can have evaluation elements from each of the major evaluation components, namely: *behavior, attitude or opinion, knowledge, outcome, administrative,* and *cost/benefit.*

EVALUATION DESIGN

Evaluation design is guided primarily by the type of questions you want to answer. However, many practical considerations influence what types of designs can and can't be used for your evaluation. In this section six basic types of designs are discussed. As you will see, each has its own unique strengths and weaknesses. No single design is appropriate for any and all evaluations. The design used must be carefully matched to the evaluation questions you want to answer and the unique characteristics of the program being evaluated, as well as other nonprogram factors which influence design selection. Each of these considerations will be discussed with each design type.

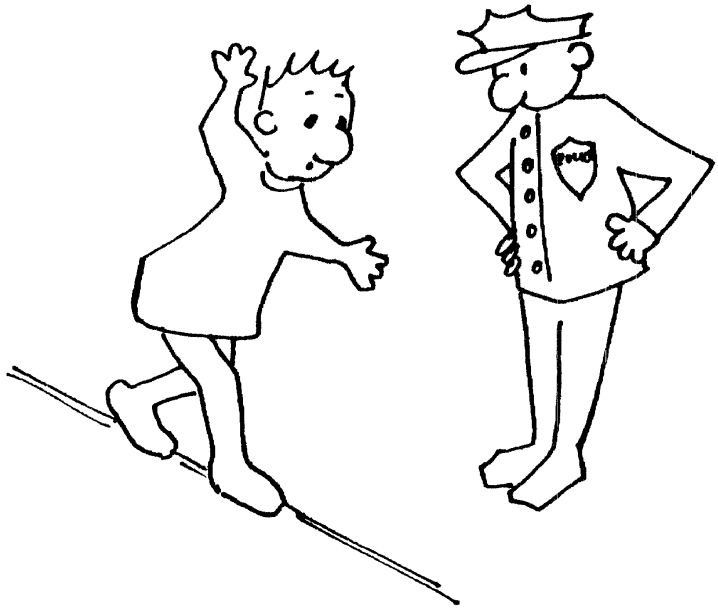
One-shot Test

The one-shot test (also called the survey design) is the simplest type of evaluation. The one-shot test involves measuring the variables of interest once at a single point in time. The best example of the use of the one-shot test is attitude, opinion, or knowledge surveys. This design can inform you about the current state of the factors of interest. A one-shot test design alone cannot provide information about the effects of a program. The design is best used when planning a program.

For example, you may want to know if there is a need to implement a safety belt use promotion program in the local high schools. A one-shot test design would involve going to the high schools, and observing safety belt use (either on one day or over several days). From these observations you can determine what percentage of the students observed currently wear belts and from that information determine whether or not a belt use promotion is necessary.

Another example, you want to know how people in your community would feel about a sobriety checklane program your department is considering. You distribute short opinion surveys to residents at local shopping centers and tabulate their responses. You find that people do not believe the drunk driving problem is sufficiently severe in the community to warrant a checklane program. You also find that people believe that a checklane program would

be an excessive intrusion into their privacy. From this information you may decide to postpone the checklane program. You may decide instead to implement a PI&E campaign to inform the public about the severity of the drunk driving problem in the community, educate residents about how checklanes are operated (hopefully reducing their concern about their intrusiveness), and thus build support for a checklane program in the future.



Cross-sectional Design

A cross-sectional design also involves collecting data at only one point in time. However, in cross-sectional designs, comparisons are made between different subdivisions (cross-sections) of the data. These subdivisions are generally two or more groups that differ in some meaningful way. Common cross-sections include male-female and age based groupings. Other cross-sections may be persons you know were previously cited for speeding vs. those not cited, or persons who annually travel more than 15,000 miles vs. those who travel less than 15,000 miles each year.

Like the one-shot test design, the cross-sectional design is most appropriate for pre-program planning. This design is not well suited for determining program effects. However, program effects can be estimated to a limited degree using the cross-sectional design. To make estimates of program effectiveness, you must be able to clearly determine subgroups that did and did not experience the intervention program. Better designs exist to use to compare effects of programs on different groups, and these are discussed later.

An example of the appropriate use of the cross-sectional design: You decide you want to implement a safety belt promotional program in your community, but are unsure how to target your limited program resources. You distribute a survey at local fast food establishments asking people about their safety belt use and general travel habits in addition to their age and sex. You break the sample into groups according to average trip length, respondent sex, and age. Your analyses find that belt use is lowest for young people, males, and people taking short (crosstown) trips. This information leads you to design a program to reach young males, focusing on promoting use at local "hangouts" for this group (e.g., fast food establishments, malls, etc.).

Pre-post Design

The pre-post design is the simplest design for evaluating program effectiveness. In this design, data are collected prior to the program and after the program is completed. The post-program data are compared to the pre-program data. Differences in the pre- versus post-program data comparisons are then attributed to the program.

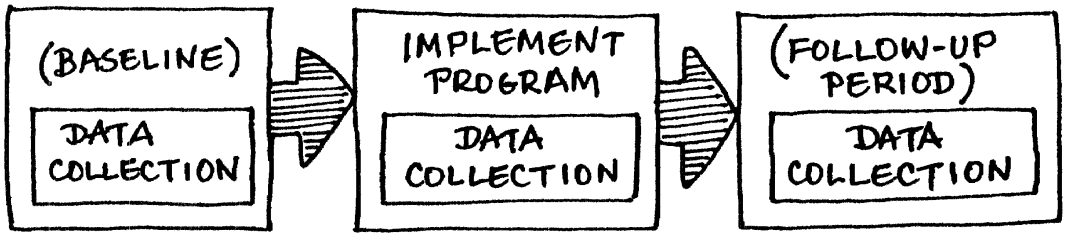
For example, the local shopping mall decides to have a traffic safety day. Among the groups participating in the mall safety day is the regional safety belt promotion council. As part of their promotion, the belt council distributes pamphlets, has informational posters, and allows people to ride "The Convincer" (a device that simulates a crash and demonstrates the usefulness of belts). To evaluate the effectiveness of their program, volunteers observe safety belt use of mall patrons as they enter and leave the mall parking lot. The data collected as they enter is the pre-program data and the data collected as they leave the lot is the post-program data. The council can be safe in claiming a successful promotion if belt use of those observed as they leave the lot is higher than belt use observed as patrons enter the lot.

This type of evaluation is most appropriate for programs conducted at a single site over a short time period (such as the program described in the example). Because of the simplicity of the design, there may be factors other than the program that are responsible for the changes observed for long-term or multisite programs. For example, some event other than the intervention program may have been responsible for the change, or the change may be part of a long-standing trend in the factor of interest. This is especially true for programs that are conducted over extended periods of time or occur over a variety of sites. Therefore, more sophisticated research designs are often preferable to the pre-post design.

Reversal Design

The reversal design gets its name from the pattern of data collection. That is, you first collect baseline (pre-program data), then collect data during the program, then collect data after the program has ended (return to baseline or reversal condition). A program's effectiveness is judged by comparing the data collected during the baseline period to the program period data and to the follow-up period data.

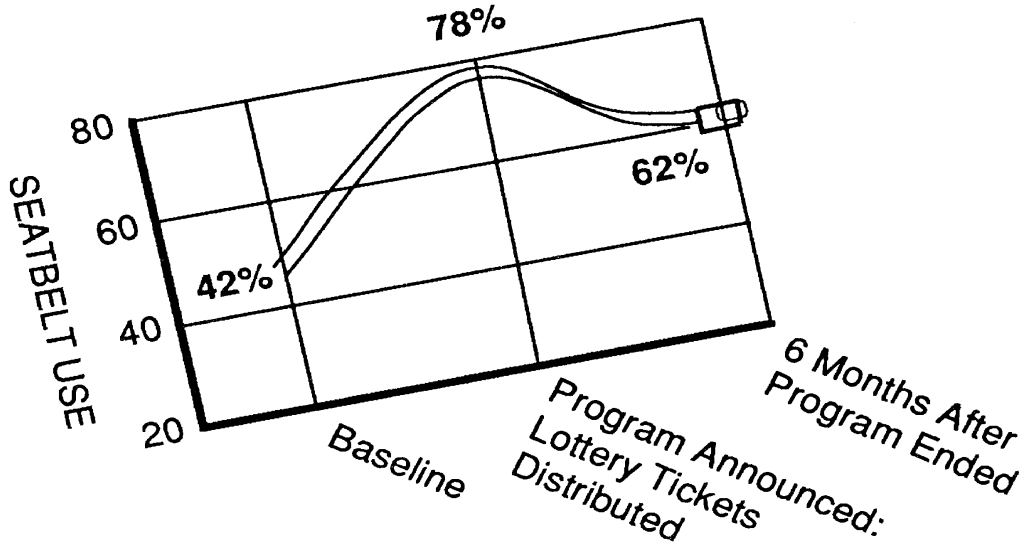
The benefit of the reversal design over the simple pre-post design is inclusion of the follow-up period as a *control* period. A control period serves as a basis to judge whether it was the program that caused the observed changes from the baseline to the program period rather than other factors coincident with the program implementation. In this design, a program would be said to have caused any observed changes if there was a change from baseline to the program period that was observed to reverse itself (return to baseline levels) after the program was ended.



Clearly many (if not most) traffic safety programs are designed to have lasting effects. The reversal design is most appropriate for programs that are expected to have short-term effects. However, use of this design will also permit you to determine what the medium- to long-term effects of your program are. For example, in many safety belt promotion programs, there is an immediate increase in belt use (over baseline conditions) due to the program. However, these increases generally decay somewhat after the intervention program has ended, but typically belt use levels do not return completely to the baseline level. In such a case, you can be confident in stating your program was effective in increasing belt use during the program, and you also have data to determine medium- to long-term effects of the program. A specific example illustrates this point.

A program's effectiveness is judged by comparing the data collected during the baseline period to the program period data and to the follow-up period data.

Acme Infant Products has decided to promote safety belt use among its employees using a lottery. Every time employees are observed using belts when they enter or leave the plant parking lot they are given a lottery ticket. Once a week, three lottery ticket stubs are drawn from the pool of tickets given out that week. Prizes are awarded to those employees whose tickets are



drawn. For one month prior to announcing the program, the plant safety department observed belt use as employees entered and left the plant and found 42% of the employees used their belts during the baseline period. When the program was announced and lottery tickets were distributed, belt use (measured by the proportion of employees receiving tickets) jumped to 78%. Six months after the program ended, and lottery tickets were no longer distributed, safety belt use was again observed in the parking lot. These observations found belt use had dropped to 62%.

From these data the safety department can conclude the lottery program was successful in promoting belt use. Belt use increased 36 percentage points during the program. The long-term effect measured by the follow-up data collection period showed that belt use levels dropped off somewhat after the program was ended, but that belt use remained 20 percentage points above the baseline level. The fact that there was a dropoff in belt use following termination of the lottery program demonstrates that it was the lottery and not some other factor coincident with the lottery program which influenced belt use. On the other hand, the follow-up data showed that the program had a positive effect on belt use even after the program had ended. These data were used to support the department's suggestion to management that the lottery promotion be reintroduced at regular intervals throughout the year.

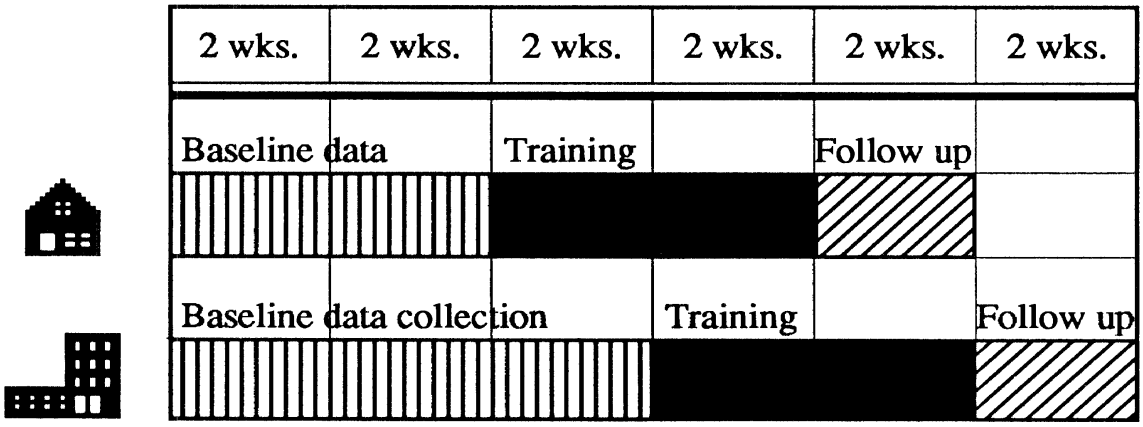
Had there been no reversal of belt use (that is, had belt use remained at 78% during the follow-up) Acme could not state unequivocally that it was the program and not some other coincident event which caused the increase. More sophisticated designs are available to better determine cause and effect relationships under these circumstances.

Multiple-baseline Design

The multiple-baseline design is quite similar to the reversal design. In fact, the multiple-baseline design is really two reversal designs being conducted simultaneously. In the multiple-baseline design, two or more sites or groups are selected for a program intervention.

Consider a program in which a one-month program of pedestrian training with elementary school children is being implemented at two separate schools in a township. First, baseline data are collected simultaneously at both schools. For example, for one month the safety patrols at both schools make observations of the number of children who cross the street at sites other than the corners.

The next phase in the multiple-baseline design is that while one school begins its training program, the other continues the baseline data collection. Thus, for a period of time there are intervention data being collected at one school and baseline data being collected at the other. For this example let's say the baseline at the second school continues for two weeks after the program began at the first school. After this two-week period of additional baseline data collection, the second school begins the training program and collects data on intervention effectiveness. For the next two weeks, both schools will have their pedestrian training programs operating. After this two-week period, the first school begins its follow-up data collection while the second school continues its training program. After the last two weeks of the training program have ended at the second school, it begins the follow-up data collection.



As the illustration shows, the central feature of the multiple-baseline design is the extended baseline period for one of the two groups. This feature could be expanded to more than two groups by extending the baseline period still further for each of the additional groups.

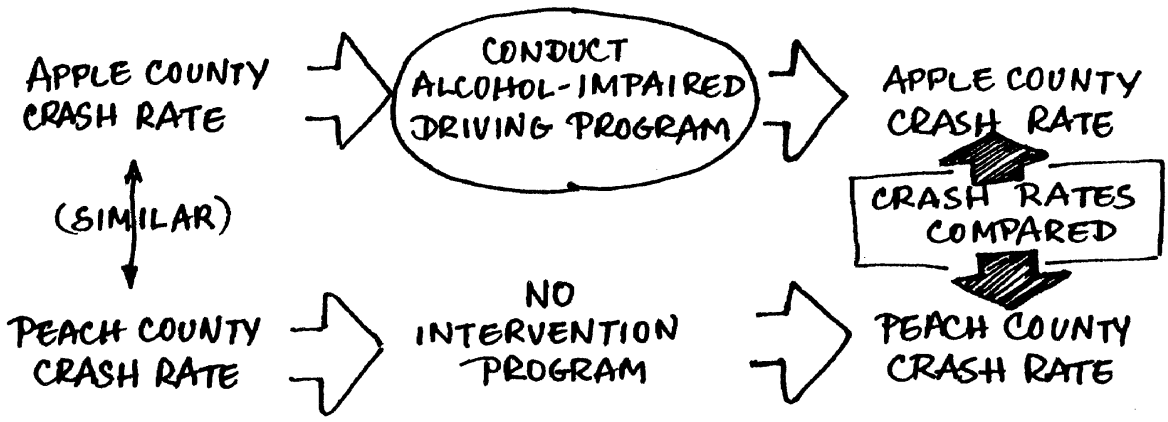
The strength of the multiple-baseline design is that you can determine whether the intervention program had an effect with greater certainty than for the reversal design. You know that the program has had an effect if the observations show a change only at the time the intervention program is in effect or is removed. On the other hand, if the data for the extended baseline observations change for the second group at the same time as the intervention for the first group is begun, then you know that either: (1) something other than the intervention is affecting both groups, or (2) the intervention program is somehow affecting both groups. To eliminate the second possibility, try to select groups in a way which minimizes the possibility of the intervention program spilling over to other groups. In the preceding example, you could try to select elementary schools which are separated geographically as much as possible to minimize interactions between the children in the schools which could contaminate program effect measures.

Control-group Design

In a control-group design, one group is selected to receive the intervention program while a second group does not receive the program. The second group serves as a *control* or comparison group. If a difference is detected between the program group and control group, the program is said to have caused the difference.

An important step in using the control-group design is selection of the control group. It is important that the control group be as similar as possible to the group which receives the program. In fact, proper selection is so important, it is described in detail in a later section of the manual.

An example: Alcohol-impaired driving crashes are found to be a problem in Apple County. Thus, Apple County police agencies begin an enforcement program increasing dedicated patrols for alcohol-impaired driving. To study



the effects of the program in Apple County, program officials select Peach County as a control group. Peach County is selected because of the similarity in the number and types of roads in the two counties, and in the previous 5 years the alcohol-impaired crash rate of Peach County has been identical to that of Apple County. In addition, Apple County officials know that Peach County will be unable to conduct any alcohol-impaired driving programs because of a lack in funds. After the one-year program in Apple County has ended, officials compare the alcohol-impaired crash rates for Apple and Peach Counties. They find that the alcohol-impaired rate for Apple County is 35% lower than that of Peach County for that year. The press release describing the results states that the dedicated enforcement program was a grand success.

The control-group design can also be combined with the pre-post or reversal designs to increase your ability to determine program effects. Take the preceding example of the Apple County enforcement program. If the Apple County officials wanted to know what the specific effect of the program was for Apple County, they might have compared crash data from the year prior to the program with data from the year the program was on-going (pre-post design). If there was a 40% decline in alcohol- involved crashes between the pre- and post-program crash rates in Apple County they might say the difference was due to their program. To be sure there was no alternate explanation, they should compare their data to data from the same two periods for Peach County. If the crash experience of Peach County is similar to that of Apple County, then this is evidence that something other than the enforcement program was acting on both counties. If there is no difference or a much smaller decline observed in the Peach County crash data, then Apple County officials can state the program was the cause of the difference between the crash experience of the two years.

Carefully consider the ramifications of withholding a program from a control group; resentment often turns into an intervention of sorts itself.

To make cause and effect statements in control group designs, it is critical that the program group and the control group be as nearly identical as possible. To the extent that they are not identical (as is generally the case in studies conducted outside a closely controlled laboratory or other controlled environ-

ment), differences observed between the two groups can be said to be due to differences between the two groups other than your program. Thus, close attention must always be paid to selection of control groups.

There are ethical and practical issues to consider in the use of control group designs. One is that use of a control group requires that a potentially useful intervention is being withheld from members of the control group. This is particularly problematic if the target of the intervention program is quite serious and the program is likely to have a positive effect. In this case, a use of a design where all groups eventually get the intervention program (like the multiple baseline design) is preferable. A second problem is that often people or groups are offended by being part of a control group. This resentment often turns into an intervention of sorts itself. You must carefully consider the ramifications of withholding a program from a control group, especially if members of the control group know they are the no-treatment control.

CHAPTER 4
Measuring Program Effects

There are many ways to collect data on factors of interest to your program and evaluation. Data collection procedures need to match the goals of the evaluation and the resources available for data collection. In this section we will describe strategies for collecting data on behaviors, knowledge and attitudes, as well as administrative, enforcement, engineering, and cost data.

BEHAVIORS

The three most common strategies for collecting data on people's behavior are self-report, observation, and physical trace analysis. Self-report measures of behavior involve simply asking people what they do or have done in the past. For example: *How often do you use a safety belt when driving? Would you say when you drive you use a belt every time, most of the time, sometimes, seldom, or never?* While self-report measures are relatively simple to collect compared to more time consuming behavioral observations, self-report data are often inaccurate.

Self-reported data are often inaccurate.

Self-report measures are dependent on the wording of the question, the desirability of various responses, and respondents' memories. Self-report items should be worded in as unambiguous a manner as possible. This means wording both the item and possible responses clearly. Possible responses must be worded so that respondents can select a response reliably. For example, it is often better to ask people to recall their behavior at specific instances than their general behavior over an extended period of time. Questions which make the respondents recall specific instances are easier to answer reliably because specific memories are recalled. General questions tap memories that are more vague, and thus require that people estimate behavior based on fuzzier general impressions of their behavior. When fuzzy memories are tapped, the chance that they are responding in a manner that is perceived to be most desirable is increased, leading to less accurate measurement.

Self-report measures that are personally or socially sensitive are most prone to error.

Self-report measures that tap issues that are personally or socially sensitive are most prone to error. For example, people may not be inclined to truthfully answer items about their alcohol-impaired driving behavior. The tendency for people to answer questions in a "socially desirable" manner can be reduced if the anonymity of the respondents can be assured. For example, we have found that self-report surveys of safety belt use result in people overreporting belt use by 9 to 20 percentage points, and that people often underreport their alcohol consumption. Self-report surveys which do not query information from which a person can be identified personally (e.g., name, social security number, driver license number) are less prone to be affected by respondents selecting responses that are perceived as being the most desirable rather than the most accurate.

Problems associated with self-report measures can be avoided by observing behavior directly rather than relying on people to self-report their behavior. If you want to measure safety belt use in your community, have volunteers stand on selected street corners and record the number of people who they see with and without safety belts on.

People change their behavior when they know they're being observed.

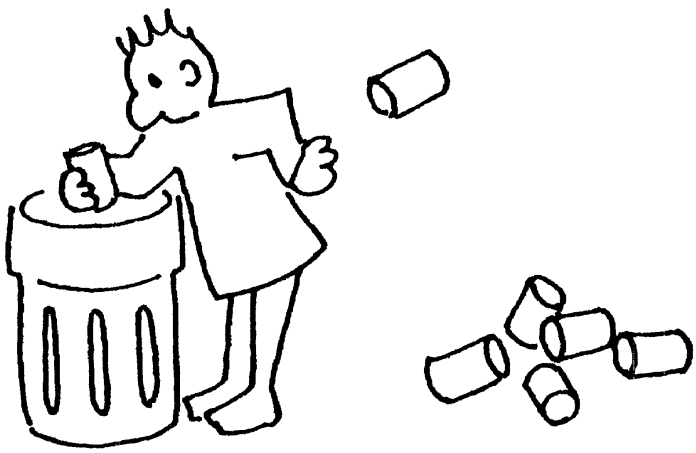
An important point when observing behavior is to remember that often people change their behavior when they know they are being observed. For example, prior to the enactment of the safety belt law in July, 1985 we observed all the cars in a stream of traffic stopped at a red light until the light turned green. Once the law was enacted, we found that people began buckling up when they noticed what we were doing. We changed our procedure to observe only three cars in the traffic stream because after observing three or more cars, people caught on to what we were doing and began buckling up.

When designing data collection forms for observations, it is important that they be as simple as possible. Observers have to be able to make their observations and accurately record what they have seen quickly. It is best if each of the possible observations (e.g., belted, not belted) be provided on the

data collection form so observers can simply put a checkmark by the appropriate category. It is also important that categories be provided for every possible occurrence (even if that means having an all-inclusive "other" category).

Observers should always check at least one response for each behavior type. This helps observers get into a routine which will increase the consistency and reliability of the data. In addition, having a code for each category will enable you to distinguish between cases where the observer forgot to code a category or the behavior was not observed. For example, if you have only a "buckled" category for drivers and it is not checked off for a given vehicle, was the driver not using the safety belt or did the observer just forget to check the box for that vehicle? You can't tell unless there was a category for "unbuckled" available. If neither category was checked, you may assume the observer forgot to code that vehicle occupant's belt use.

It is important to remember that missing data should be treated differently than data indicating the behavior was not observed. When data on a particular behavior is missing, you cannot state what the person was or was not doing. Only when you have data on the specific occurrence or nonoccurrence of a given behavior can you state what occurred. Thus, when calculating the percentage of persons observed using safety belts for example, you should calculate the percentage of persons observed using belts divided by the total number of *valid* observations. That is, divide the number of persons observed using belts by the total number of observations minus the number of observations in which belt use information is missing.



Behavior can also be measured by using physical traces of the behavior's occurrence. Although this sounds complex, it is really quite simple. Take the example of trying to measure the beer drinking behavior of persons in a park to evaluate a new prohibition on beer drinking as part of a drunk driving reduction program. You could try to observe people drinking beer at the park directly by traveling through the park trying to see people with beers in their hand or at their table. You may or may not be able to see them actually drinking, and you would have to constantly travel through the park to get a thorough count of drinkers. You could almost certainly never get an estimate of how much beer was consumed

using this procedure. Physical trace analysis of beer drinking behavior in this case could involve waiting until the park closes and counting the number of beer containers disposed of in the park.

Let's try an example less messy than garbage picking. You are interested in determining pedestrian trafficways through an grassy area where a number of roads intersect to place pathways and traffic control devices to reduce motor vehicle-pedestrian crashes. You don't need to directly observe or videotape the area to determine what paths pedestrians use most often. By examining the amount of wear in the grass where pedestrians walk you can determine the most highly traveled pedestrian paths, and base programmatic changes on this physical trace information. In fact, prior to laying sidewalks across open spaces, many groups allow pedestrians to walk where they please and lay sidewalks and traffic control devices along the most traveled paths. This way pedestrian traffic can be controlled better than laying paths first, just to have pedestrians take shortcuts across areas where paths and appropriate traffic control devices are not in place.

KNOWLEDGE, ATTITUDES, AND OPINIONS

Knowledge, attitudes, and opinions are all measured using the same basic strategies: questionnaires (surveys or tests), interviews, and group discussion. Each of these data collection methods is too complex to detail fully in this manual; however, the fundamentals of each will be described.

All of us have filled out a questionnaire at some time in our lives. There are two basic types of questionnaire items: multiple response and open ended. Multiple-response items provide the respondent with a set of choices to select from. An example is: *Compared to car drivers, do you think drivers of semi-trailer trucks drive more safely, less safely or about equally safely?* The respondent is asked to select one choice from the list. Open-ended items have no list of response options to select from. For example: *What are the three worst traffic safety problems confronting drivers in Apple County?* In open-ended items, respondents must provide their own unique answers. Each type of item has significant benefits and drawbacks.

In open-ended items, respondents provide their own unique answers.

Multiple-response Items

The strength of multiple response items is that they are time efficient

Multi-response items are useful when you have specific options you want respondents to consider.

(responses are given in the item itself), data can be easily coded (each response is predetermined and can be assigned a value for coding), and there are specific response options you want respondents to consider. The weakness of multiple response items is that respondents may not agree with or want to choose any of the response options given, and multiple response items are more difficult to write because the response options and question must both be written to reduce ambiguity, but at the same time not lead respondents to select any specific option.

Item wording can have an important impact on the responses people are likely to give. The item must provide enough information so respondents can provide a meaningful and informed answer, but the information should not be presented in a way that leads people to any given response. For example, you may want to know people's opinions of the use of sobriety checklanes as a strategy to reduce drink driving in your area. A poorly worded items would be: *Drunk driving is a leading cause of death and injury on our roads. Would you support using sobriety checklanes to reduce the number of deaths and injuries caused by drunk drivers?* There are several problems with this item.

First, people might not know what sobriety checklanes are. While there may have been much publicity on checklanes in your area, you need to establish a common understanding of the issue for each of the respondents. Use a brief definition of the issue early in the item. *Sobriety checklanes are police patrols where all drivers on a given road are briefly stopped to determine if they are driving while impaired by alcohol.*

Respondents may not agree with <u>any</u> of the response options given.

Second, it frames the issue towards an affirmative response by providing information about the effects of drunk driving. While we would often like to know how people feel about an issue when they are informed about the facts surrounding the issue, many people may not know those facts. Therefore, you should not provide background information which supports or undermines a particular response choice. A more neutral wording would be: *Sobriety checklanes are police patrols where all drivers on a given road are briefly*

stopped to determine if they are driving while impaired by alcohol. Some have suggested that sobriety checklanes are an effective way to reduce the number of drunk driving crashes. Others dispute their effectiveness as an unnecessary intrusion on personal liberties.

Third, there is only one response item described in the question (i.e., *Would you support. . .*) In addition, the wording in the response given is biased toward an affirmative response. Most people would support a program that is almost promised to reduce deaths and injuries. All possible response options should be given in the question, and the response wording should be kept neutral. While this creates items that often seem quite wordy, it reduces the bias toward or against any of the possible response options. A better wording would be: *Sobriety checklanes are police patrols where all drivers on a given road are briefly stopped to determine if they are driving while impaired by alcohol. Some have suggested that sobriety checklanes are an effective way to reduce the number of drunk driving crashes. Others dispute their effectiveness as an unnecessary intrusion on personal liberties. Would you favor or oppose use of sobriety checklanes as a strategy to reduce drunk driving crashes in your community?* The use of "as a strategy to reduce. . ." changes the possible bias introduced by the original wording "to reduce the number of deaths. . . "

People often select the item in the middle of the spectrum of choices.

Multiple response items also are prone to a phenomenon called "central tendency." Central tendency means that people very often select the item options that are in the middle of the spectrum of choices. For example: *If a person is not using a safety belt and is stopped for speeding, how likely is it they will get a ticket for not having a safety belt on? Would you say there is almost no chance they would get a ticket, it is unlikely but it happens sometimes, there is a good chance of a ticket, they will get a ticket nearly every time, or they will always get a ticket?* In a recent survey of Michigan residents, responses to this item most frequently fell in the middle category (there is a good chance of a ticket) despite the fact that the chance of being ticketed is quite small. It is difficult to overcome central tendency effects, but omitting response options that are wishy-washy helps. Omit options like "neither agree nor disagree" if you want to force respondents to make a choice between agreement or disagreement to an item.

Open-ended items are well suited to provide information where you are unsure what specific responses are appropriate. In fact, these types of items

are often used to provide information for the development of multiple response categories for future surveys. Open-ended items are also useful if for some reason you don't wish to prompt the respondent with possible answers to your question. Of course, you must always be sure that the wording of such open-ended items does not lead the respondent to any specific response set.

If you want a rich source of data, consider using open-ended items.

Open-ended Items

Open-ended items generally provide what is called "rich data." That means that the detailed answers that respondents give to open-ended items generally contain more information about the specific issue queried than do multiple response items. If you want a rich source of data, you should consider using open-ended items.

While open-ended items provide rich data, coding, analyzing, and interpreting these data is quite difficult. While the richness of narrative responses is useful in getting a better understanding of the issue as a whole, you are also likely to want to make specific statements like, "43% of the sample reported. . ." You must go through each respondent's answers and determine how to categorize their responses if you want to make use of the group data in aggregate. You will probably find it quite difficult to boil down responses from open-ended items into concise, meaningful data. Open-ended items are also quite time consuming to use in surveys. It takes people much longer to give a narrative response to an item than to pick a response category. This time and effort is compounded if the survey is being conducted via telephone because phone survey staff have to type or handwrite the responses.

This brings us to the differences between questionnaire surveys and interviews. Often questionnaires and interviews differ only in the form the presentation of the material takes. People generally fill out written questionnaires on their own, while some interviews are simply oral versions of the written questionnaire. Important distinctions should be made between questionnaires that are written and those which are presented orally. Because the mode of presentation differs, the form of the information should be tailored to the presentation mode.

Points that are easily illustrated on a written form are often more difficult to describe orally.

Written surveys may ask respondents to refer to earlier portions of the survey, and the respondent could just turn the pages back to that section. In telephone surveys, this would require the respondent to remember what had happened previously. It is not advisable to count on respondents' memories. When using an oral survey format, make every point clear at each stage of the survey. Also remember, points which can be easily illustrated on a written form are often more difficult to describe orally.

Interviews may be less structured than a survey presented orally. Semi-structured interviews are simply discussions with individuals about topics the interviewer selects. The interviewer has a set of general questions he or she uses to guide the discussion, but the interviewer may let the interviewee guide the specific course of the interview.

Similarly, group discussions can be used to elicit information from a number of people at the same time. Typically group discussions involve 4-8 people sitting around a table with a discussion moderator whose job it is to ensure that the discussion stays on the general topics at hand, and that no single person or small group of people from the larger group dominates the discussion. While interviews and group discussions are often audio and occasionally videotaped, another critical job for the interviewer or discussion moderator is to take detailed notes of the sessions.

Semi-structured interviews or group discussion sessions often provide very rich data, but these data are very difficult to quantify. The techniques are best used when you want to know how a person (or a group of people) feel about a range issues but are unsure about what specific questions to raise. Often, group discussions and personal interviews elicit information that is difficult if not impossible to elicit from written surveys. The real plus of personal interviews and group discussions is that the questions used to elicit responses can be modified by the interviewer or group discussion moderator depending on the current discussion.

The moderator must be able to listen and facilitate discussion without interjecting personal opinions or bias into the session.

The skills necessary for a person acting as an interviewer or group discussion moderator are good active listening and interpersonal skills. An interviewer or moderator must be able to listen and facilitate discussion without interjecting personal opinions or bias to the session. Active listening skills include letting the person or group know you are

following what they are saying by repeating it back to them and querying them for further thoughts on the discussion subject. The use of phrases such as *I understand,* or *What you're saying is...* followed by a brief summary of one or more of the interviewee's recent statements, or *Could you expand on that?* keeps the flow of information going by letting the interviewee know you are listening and interested in what they have to say. In group discussions your key role will be to keep the discussion on track and to make sure that no one person dominates the discussion. You should elicit responses from specific individuals who seem to be less active in the discussion, and reinforce their participation with words of support and encouragement.

ADMINISTRATIVE DATA

It is important that you collect data on the administration of your project. This is especially true if the project is primarily administrative (e.g., setting up a traffic safety task force for your community). While straightforward, administrative data collection is often overlooked. If your program involves distributing information pamphlets, collect data on how many pamphlets are distributed, where they are distributed, when they are distributed, and what groups (or individuals) distribute them. Record information on media contacts, and stories that are written or broadcast about your project. Collect information on who you asked to attend various meetings, who actually attended, and briefly describe each member's participation. Record the amount of time project staff spends on various tasks for the project.

POLICE AND TRAFFIC ENGINEERING DATA

A variety of data are regularly recorded by police and traffic engineering agencies that are useful for program evaluation. These data include data on traffic crashes, traffic citations, traffic speeds, and traffic volumes (the number of vehicles traveling on a given stretch of road during a specific time period). It is important to work with your local, county, and state departments of transportation, road commissions, and police agencies to determine what data are already available and what data can be collected to assist with your evaluations.

Discuss your data needs with these organizations during the development of your program evaluation plan.

Once you determine what data are available, it is important to have persons who are knowledgeable about the way the data are collected and the specific meanings of the data codes to help you

interpret the data. It is often the case that seemingly simple data are misinterpreted by users unfamiliar with the nuances of the data collection and coding schemes. It is best to discuss your data needs with these organizations during the *development* of your program evaluation plan. Remember, traffic safety is an important part of their jobs. You will find these organizations are almost always as interested in knowing the effects of your program as you are. Often these agencies can help you collect data you may not have known were even available. Work closely with them and your evaluation efforts will undoubtedly be strengthened.

COST DATA

One of the most often overlooked aspects of project evaluation is cost. Cost evaluations are generally considered to fall into one of two categories: cost-benefit and cost-effectiveness. Simply stated, cost-benefit analysis is used when comparing the costs of a given program to the benefits derived from that program. For example, a program cost $20,000 and benefits derived from that program are valued at $30,000. You would then state that there was a positive cost-benefit ratio for the program (i.e., the value of the benefits from the program exceeded program costs) and that similar programs should continue to be enacted.

Cost-effectiveness evaluations are used when comparing costs and benefits of several similar programs. The purpose of cost-effectiveness evaluations is to determine what programs provide the most benefit for the least cost. Consider three programs (Table I): program A cost $10,000 and returned $15,000 in benefits, program B cost $20,000 and returned $25,000 in benefits, and program C cost $30,000 and returned $50,000 in benefits. Program C returned $1.67 in benefits for every dollar invested, program A returned $1.50 for every dollar invested, and program B returned $1.25 for every dollar invested. This information helps decision makers select what programs to implement with available resources.

Table I
Cost Effectiveness of Evaluations

	Cost	Return	Cost Effectiveness
Program A	$10,000	$15,000	$1.67
Program B	20,000	25,000	1.50
Program C	30,000	50,000	1.25

Cost data should be collected regularly. These costs should include more than "out-of-pocket" expenses for salaries and materials. Try to establish values for items which have been donated to the program. These may include donated materials and services, prizes or incentives you may distribute, ad space in newspapers and broadcast media, and volunteer staff time. In your evaluation, include reports of both actual "out-of-pocket" costs and the value of donated materials. Funding agencies will be impressed by the value of services and materials that have been provided at no cost to the project, and others interested in conducting similar programs will know how much the program would cost if they are unable to obtain similar donations.

It is also valuable to consider separating the costs for conducting the program itself from the costs associated with evaluating the program. For example, if you are sponsoring bicycle helmet use promotions in the local school district, collect data on the cost of materials and manpower associated with both the program itself and the evaluation of the program. In this way, you will be able to know the cost of the program independent of the evaluation component. Both sets of figures are valuable for assessing the total cost of the program.

One of the most difficult components of cost evaluations is determining the value of deaths and injuries prevented by a traffic safety program. There are direct costs (i.e., property loss and damage, direct medical and mental health costs, emergency services, administrative costs, and productivity loss), and losses (costs) associated with reductions in the quality-of-life a person would have enjoyed had that person not been injured. While there are a variety of ways to describe costs associated with traffic crash injuries, perhaps the most simple is to describe costs in terms of injury severity levels which are currently used on Michigan crash reports (the UD-10 form).

Crash reports in Michigan use the KABCO injury scale. In this scale there are 5 levels of injury:

K— the person died within 30 days after the crash;

A— severe or incapacitating injuries which prevent injured persons from walking, driving, or normally continuing activities they were capable of prior to the crash;

B— moderate or nonincapacitating injuries evident to observers at the scene of the crash;

C— minor or possible injuries reported or claimed which are not fatal, incapacitating, or nonincapacitating; and

O— incidents in which no one was injured resulting in property damage only.

Table II
Costs of Traffic Crashes (Per Incident)

Injury Severity	Direct Costs	Quality-of-life Costs	Total Costs
Fatal	$426,272	$1,602,710	$2,028,982
A-level	16,251	51,032	67,283
B-level	4,177	9,598	13,775
C-level	2,761	4,270	7,031
Property damage only	1,415	227	1,642

Table II provides costs associated with each of the KABCO levels of injury in terms of direct costs, quality-of-life costs, and total costs. For the purpose of a cost evaluation, we recommend use of the total costs. These costs best describe the entire value of loss of life and injury caused by motor vehicle crashes.

CHAPTER 5
Sampling

Sampling is used to obtain data on a subset of a group so that inferences about that group can be made based on the data from the sample. It is generally easier to obtain information from a subset than from the entire group. Consider the case of safety belt observations. We are unlikely to be able to observe every person in a community to determine their safety belt use or make observations at every intersection within a community, even with enormous resources and patience. However, we can sample people or locations to make inferences about belt use in the community at large. To select a sample strategy we must know to what set of people, events, locations the study's conclusions are intended to apply. There are a variety of sampling strategies that can be applied to collect representative data.

RANDOM SAMPLING

A random sample has the following characteristics: (1) every member of the group is available to be included in the sample, and (2) every member has an equal opportunity to be included in the sample. Using the safety belt observation example, let's suppose we want to select intersections at which to conduct the observations. One way to get a random sample of intersections would be to put every intersection name on a slip of paper, place these slips in a bowl, mix the bowl thoroughly, and draw out slips of paper until the desired number of intersections has been selected.

The main strength of a random sample is that it is representative of the larger group from which it is taken.

At this point you may be wondering why an evaluator would want to go to so much trouble in selecting observation sites. Why not just go out to a few "key intersections." While going to a few "key intersections" is a valid sampling strategy in its own right (see purposive sampling), it provides different information than a random sample. The main strength of a random sample is that the sample is representative of the larger group from which it

is taken. We generally don't know whether or not the sample truly has the same characteristics as the larger group, but with random sampling we know that the probability of our sample being representative is greater than that of other sampling strategies.

SYSTEMATIC SAMPLING

In a systematic sample design, a list of all the larger group elements is made (e.g., list all intersections). A random starting point on the list is chosen after which every Nth intersection on the list is chosen. The value of N in this case is determined by dividing the number of intersections by the total number of intersections you want included in your sample. To maximize the representativeness of your sample to the larger group, intersections (or other sample units) should be assigned to the list in a random sequence.

STRATIFIED SAMPLING

A stratified sample design is typically a two-stage process. First, the larger group is divided into strata using definitions that permit groupings into mutually exclusive categories. That is, categories which no person or thing can be considered a member of more than one category. For example, you could stratify intersections based on the amount of traffic which the intersections carry (often available from the department of transportation or local road commission). In the second stage of the sampling procedure samples are randomly selected from each of the strata.

QUOTA SAMPLING

Quota sampling involves selecting predetermined numbers of people or things from each of several distinctive subsets of the larger group. Consider the following example. You want to determine what factors are related to drivers' speeding behavior. You believe that age and sex are likely to be important correlates of speeding. Rather than getting a random sample of drivers to interview about attitudes about speeding and previous speeding behavior, you decide to interview 20 randomly selected persons in each of 4 groups: males under age 25, females under age 25, males age 25 and older, and females age 25 and older. Quota sampling is typically the

Quota sampling ensures that you sample a sufficient number of people or things in each of your desired categories.

preferred sampling strategy when you want to ensure that you sample a sufficient number of people or things in each of your desired categories. A purely random sample cannot assure you any specific makeup of subjects. In fact, the distribution of subjects in your sample will generally reflect the general distribution in the larger group from which the sample is drawn.

PURPOSIVE SAMPLING

Purposive sampling assumes that the evaluator applies some sound judgement to hand pick people or things from the larger group for the study. For example, you may wish to determine the administrative opportunities and obstacles that may affect a law enforcement program to reduce alcohol-impaired driving. A purposive sampling strategy might involve interviewing "key players" in the community who can affect the program's success or failure. Thus you may wish to interview the police chief, sheriff, alcohol rehabilitation coordinator of the local hospital, a "key" city council member, a judge and prosecutor likely to handle the cases, and a lawyer from the local defense bar.

Purposive sampling should only be used in cases where the representativeness of the study's results is not a concern.

You can see that purposive sampling does not involve random selection of people or things from the larger group. Therefore, results from a purposive sample design cannot be generalized to any larger group. Purposive sampling should only be used in cases where the representativeness of the study's results is not a concern.

CONVENIENCE SAMPLING

A convenience sample is a nonrandom sample in which data are collected from people or things from a larger group because they are easily available. For example, police officers may be encouraged to observe safety belt use while on patrol when not busy with other duties. While you cannot generally generalize results from a convenience sample to the larger group, often the choice in evaluations is to either use a convenience sample or not to collect any data at all. The obvious choice between the two is generally the convenience sample. Data from convenience samples are more compelling when several studies using convenience samples result in similar findings. As with each of the sample designs described here, you must consider the nature of the sample design when discussing your results.

CHAPTER 6
Data Analysis

While data analysis usually is conducted after all the data have been collected, it should really be considered and planned prior to the first data collection session. How you plan to analyze your data will have a major impact on the research design as well as the data collection strategies you use. Data analysis techniques range from simple graphing of the data to extremely complex mathematical modeling procedures. As we will describe, there are generally valuable resources in your area that you can tap for assistance with the more complex analysis strategies.

GRAPHIC ANALYSIS

Almost any data analysis plan should begin by graphing or charting the data. Basic graphic analysis is usually in the form of pie charts (Figure 1), bar charts (Figure 2), or line charts (Figure 3). While you can inspect the numbers you obtain from your data collection procedures without somehow graphing them, simple charts often highlight features of the data which may not be noticed otherwise.

Pie charts are appropriate when you want to show what proportion of a whole each subgroup comprises. Figure 1 shows the proportions of all respondents to a survey who selected each of the response items. This figure shows that 54% of the respondents report they believe freeways in Michigan are in

Figure 1
Condition of Michigan's Freeways

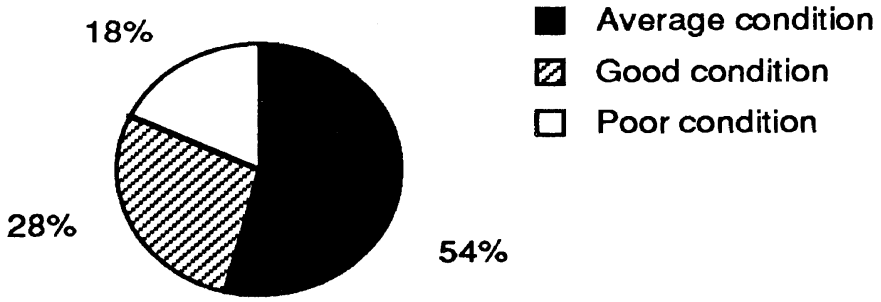
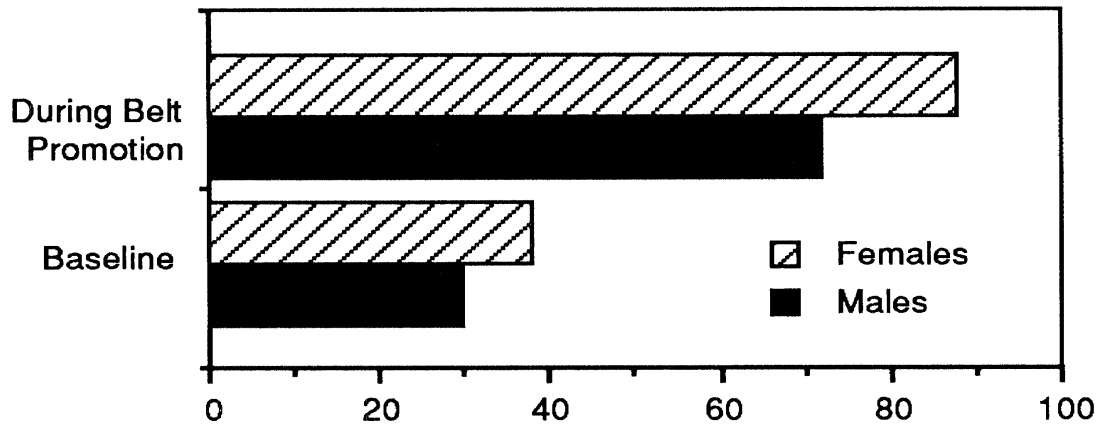


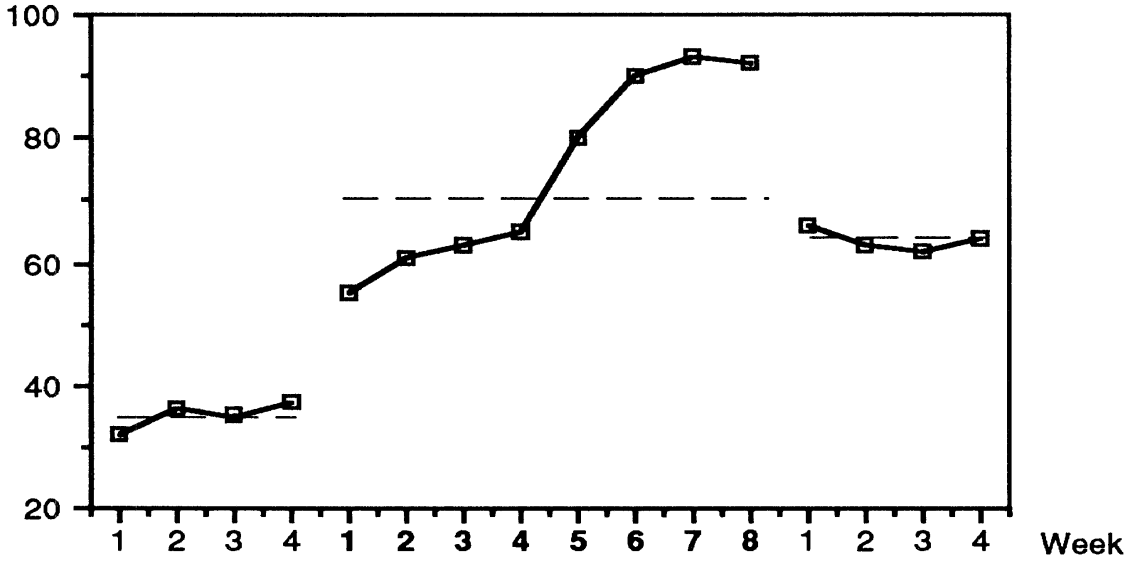
Figure 2
Observed Belt Use at Hawthorn High



average condition, 28% believe freeways are in good condition, and 18% believe freeways are in poor condition. Pie charts are most appropriate when describing levels of a single measure or variable.

Bar charts are best when you want to compare results of a given measure across a limited number of groups (two variables). The example in Figure 2 shows the proportion of males and females who were observed using belts during the baseline phase of the high school belt promotion. The chart shows

Figure 2
Safety Belt Use at Hemingway High



that belt use was lower among the males than the females. Bar charts are most effective when comparing two variables (in this case belt use and sex). Three-dimensional charts are often used to show relationships between three variables (e.g. belt use by sex and high school class), but the charts become more difficult to interpret because of the amount of data shown in one chart.

Line charts are particularly well suited to illustrate trends in data across time (as one would have in several or multiple baseline designs). Figure 3 shows a possible line chart for safety belt use collected at a high school before, during, and after implementing a safety belt promotion program. As you can see from the chart, belt use increased from the baseline level during the promotion phase then declined slightly after the program was completed.

STATISTICAL DATA ANALYSIS

You will usually want to be able to answer questions like: *Did my program cause any change in. . . ?* or *Are there differences on this measure between Group A and Group B?* Analysis of your data using some sort of mathematical or statistical procedure is the most common way to answer these types of questions. Statistical analysis determines the probability that program effects are "real" and not simply artifacts caused by the research methods used to collect the data.

Many of you have probably heard the term "statistical significance." Researchers are often asked the question: *Are these differences statistically significant?* What the question generally implies is: *Are these differences real?* The issue of statistical significance is most important when using a design where samples are taken to represent a larger group. Statistical significance refers to whether or not differences observed in data collected from samples are due to "true" differences between comparison groups or are due instead to errors that are inherent in sampling and measurement methods. The only way that samples are error free (that is they represent the larger group from which they are selected) is when the sample size is equal to the size of the larger group. The closer the sample size is to the size of the larger group, the less "sample error" that exists.

The closer the sample size is to the size of the larger group, the less "sample error" that exists.

The behavior, knowledge, and attitudes of individuals in a group are seldom

identical to the "average" behavior, knowledge, and attitude of the group taken as a whole. Statistical significance also considers how much the data varies around a given central point. When data vary a great deal around a central point (such as the arithmetic mean), determining differences between two means becomes more difficult.

The mean is the arithmetic average of the data. To calculate the mean you add up all of the data points and divide by the number of data points. For example, you measure safety belt use on 4 occasions during baseline (observing the same number of drivers each day) and find belt use on those days was 43%, 47%, 45%, and 42%. You add up these data (177) and divide by 4 and find that mean safety belt use during baseline was 44.25%. You continue to collect data during the program, and on the four observation days you find belt use to be 47%, 60% 65%, and 68% (mean=60%).

Notice that on none of the baseline days was belt use observed to be 44.25%, and on only one of the intervention days was belt use actually 60%. That is because average belt use rates (like almost every behavioral phenomenon) varied around the mean. The question of statistical significance is: *Is there so much variance around these two means that there is no "real" difference between them?* There are a wide variety of statistical procedures available to answer this question. Each statistical procedure is appropriate for a given type of data, the hypothesis being tested, and research design.

For our purposes here, it is sufficient to be introduced to the idea of statistical significance. Actually analyzing data using most statistical tests requires more training than can be provided in a brief manual such as this. However, you should consider consulting with persons in your local community who may have a better understanding of statistical analysis than yourself. Often local community colleges have staff who are sufficiently versed in statistical analysis to be of help. For the most part, you want to stay away from research protocols that require highly complex analyses. Results from the more complex analyses are often difficult to interpret clearly and should be avoided for most community program evaluations.

ANALYZING CRASH DATA

Many of you at one time or another will want to answer the question: *Did my program reduce the number of crashes in the community?* or *Did our program reduce the number of injuries or deaths resulting from vehicle crashes?* Unfortunately, these are the most difficult questions to answer in a traffic safety program evaluation. This is especially true for evaluations of relatively small geographic areas such as cities or counties. In fact, it is often difficult

to analyze crash data for an area as large as the entire state.

Probably the most serious problem facing evaluations of crash data is that crashes (especially those resulting in injury or death) are relatively infrequent events. Crashes have highly seasonal cycles which influence any potential evaluation, particularly in northern states like Michigan. In general, there are fewer crashes in the summer than the winter. On the other hand, there are more injury and fatal crashes in the summer than in the winter. Summer crashes as a whole are more severe than winter sliding, fender-bender crashes.

That leaves us with the question of how to evaluate program effects on crashes and crash injuries. The most simple (but unsatisfying) answer is: don't.

That leaves us with the question of how to evaluate program effects on crashes and crash injuries. The most simple (but unsatisfying) answer is: don't. Focus your evaluation on behavior or attitude changes rather than the ultimate outcome of crash or injury reductions. If you are promoting safety belt use (with the hopeful outcome being reduced injuries in crashes), evaluate the effects of the program on safety belt use. If you are conducting an education program on the laws regarding stopping for school buses, measure knowledge and attitude effects. For such a program, you may also want to try to evaluate behavior change of drivers around school buses.

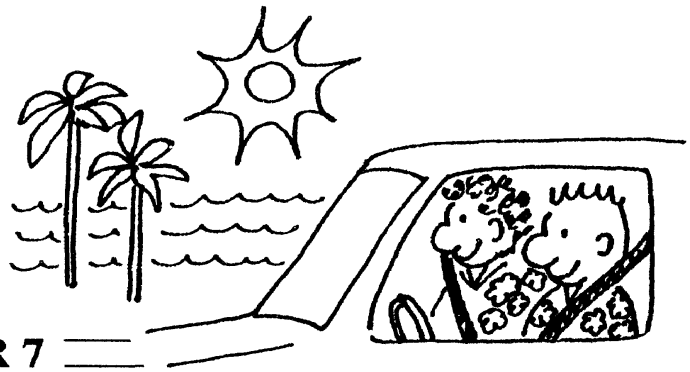
If you decide you want to include a study of crashes or crash injuries in your evaluation, here are a few tips. First, gather as much data on the time period preceding your intervention as possible (we're talking years here). If possible, compare data from similar seasons. For example, if you have a speed reduction program being conducted by the police in the summer months, compare speed-related crashes for those months compared to those same months for the previous 4-5 years.

Remember that crashes are related to the amount of driving people do. That is, as the amount of travel increases, so does the number of crashes. Two measures of travel are often available from the department of transportation or road commission: average daily annual travel (traffic counts from specific roads) and vehicle miles of travel. Aside from using highly complex analytic techniques, your crash data analyses will be strengthened by comparing the number of crashes per mile traveled or crashes per number of vehicles

traveling on the road. That is, use the rate calculated by dividing the number of crashes by the vehicle miles traveled or traffic count. If you don't consider travel factors, findings of a seemingly effective program (i.e., one that seemed to reduce the number of crashes) may in fact be the result of a reduction in travel, not an effect of your program. Conversely, a program may seem to have had no effect until you analyze the rates per mile traveled and find that although the total number of crashes has increased, the rate of crashes per mile traveled has decreased.

Remember that crashes are related to the amount of driving people do. That is, as the amount of travel increases, so does the number of crashes.

Similarly, some programs are not designed to reduce the number of crashes, but rather the number of injuries and deaths resulting from crashes. When analyzing crash outcomes such as injuries and deaths caused in crashes, it is important to consider the rate of injuries or deaths per crash. In this way you are controlling for the possibility that an apparent decline in deaths is really due to a reduction in crashes rather than your program. A perfect example is a safety belt promotion program. One does not generally expect safety belt use to reduce the number of crashes that occur (with some notable exceptions). However, increased belt use should reduce the number and severity of injuries given a crash occurs. Thus, you should compare the rate of injuries or deaths per crash before the program to the same rate after the law. This rate is calculated by simply dividing the number of injuries or deaths by the total number of crashes.



CHAPTER 7

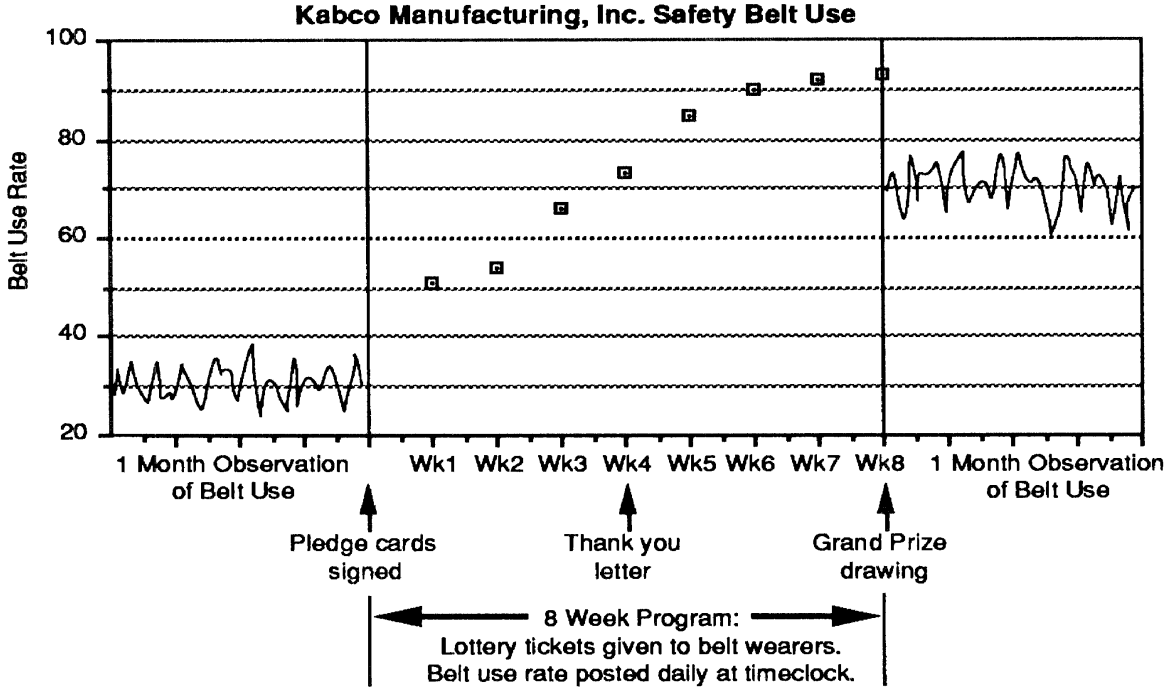
Designing an Evaluation Plan: Pulling It All Together

In this chapter we have presented several strategies for evaluation design, data collection, sampling, and analysis. You may find the wide variety of alternatives helpful but confusing. What design should be used for your program evaluation? It's a difficult question with few absolute right and wrong answers. In this section we will try to pull all this information together with three examples of programs and possible evaluation plans. The three examples we will discuss are: (1) a corporate safety belt program at a single work site, (2) a bicycle helmet promotion effort at an elementary school district, and (3) a police enforcement program to reduce speeds and increase belt use on a stretch of highway in Apple County.

Our first example is a corporate safety belt promotion. Kabco Manufacturing, Inc. conducted a *one-shot test* to measure employee safety belt use as employees enter and leave the plant parking lot. The plant safety staff conducted one week of observations and found belt use at the plant to be 34% (below the recently announced state average of 49%). Because of this low rate of belt use, Kabco met with appropriate plant personnel to discuss options for increasing belt use (shop stewards, safety staff, and management). This committee decided a program to promote safety belt use among its employees was a good idea.

The program the committee decided on (Figure 4) involved having employees sign "pledge cards" that the employees hung on their rear-view mirrors promising safety belt use for vehicle occupants. Employees signing pledge cards were entered in 8 weekly lotteries for prizes donated by local merchants (a two-month program). During the two-month program period, employees displaying pledge cards on their mirrors received additional lottery chances every time they were observed wearing their belt as they entered or left the Kabco parking lot. In addition, the program included a special incentive clause. If the safety belt use rate for employees reached 90% by the end of the two-month period, there would be a special prize drawing for a 2-week vacation in Florida (donated by a local travel agency). Employees would receive one chance to win the "big prize" for every time they were observed buckled up by the safety staff.

Figure 4
Kabco Manufacturing, Inc. Safety Belt Use



To evaluate their program, Kabco selected a *reversal* design. A reversal design was selected for the evaluation because the program would be conducted at only one site, and the program itself would be only two months long. Therefore, there was an opportunity to collect belt use data prior to the program, during the program, and after the program ended. A *control group* design and a *multiple baseline* design were both rejected as possible evaluation plans because there was no alternate site to serve as a control site or to conduct a parallel program. A *pre-post* design was rejected because project staff wanted to collect information about short- term follow-up effects after the program had ended.

Kabco safety staff conducted belt use observations of employees as they entered and left the employee parking lot. During the program (intervention) phase, observations also included distribution of lottery tickets to employees observed wearing their belts. To inform employees of company-wide belt use, the safety staff posted a chart of daily belt use at the time clock where employees punch in and out which was updated each day. One month after the end of the lottery phase, safety staff conducted one month of observations to determine short-term follow-up effects of the program.

During the baseline and follow-up data collection periods, observers remained "unobtrusive" by conducting their observations from a car parked near the parking lot entrance. This was done so the belt use of employees

entering and leaving would not be affected by the knowledge that they were being observed. During the lottery program, safety staff stood by the gate in orange safety vests so they could easily observe belt use and distribute lottery tickets to persons observed wearing a safety belt.

Prior to announcing the program, the safety staff conducted one month of daily safety belt use observations as employees entered and exited the plant parking lot. The results of these observations were included in the information packet which described the planned program and included the pledge cards.

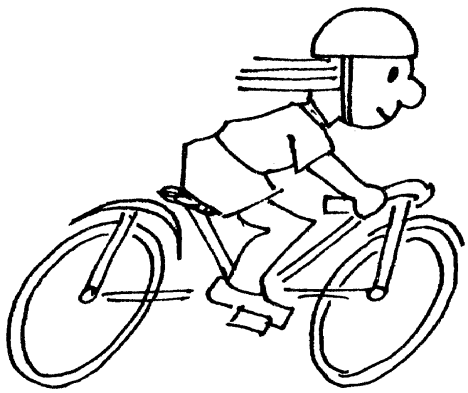
During the two-month program period belt use was observed daily, and the employee belt use rate was plotted each day on the chart at the time clock. At the halfway point, a memo went out to employees thanking them for their participation and increased belt use (a chart of daily belt use to date was included in the memo). The memo also included the names of the first four lottery winners, and informed employees that belt use would have to increase if they wanted to qualify for the grand prize drawing. Safety staff continued to observe belt use and distribute lottery tickets for the next 4 weeks. As you can see in Figure 3, belt use reached 90% in the sixth week of the program period, and remained at or above 90% until the grand prize drawing. One month after the grand prize drawing the safety staff began a 4-week follow-up observation period. Safety belt use had declined from the program high of over 90%, but remained well above the baseline belt use average of 35%.

Cost data were also collected for the evaluation. Because all of the lottery prizes were donated by local merchants, there was no cost to the company. However, the retail value of the prizes was determined to provide information on what prize costs would be if donations could not be solicited for future programs. Safety staff spent 2 hours each day of the program observing belt use and distributing lottery tickets. During the course of the program, 2 employees were involved in traffic crashes. Both reported they did not use safety belts prior to the program and that they began to use belts because of the program. Neither employee was injured in the crashes. Police officers who investigated the crashes reported that both employees would probably have received substantial injuries had they not been buckled up. Further investigation showed that each employee would probably have lost 3-8 work days recovering from injuries prevented by the safety belts. Cost savings for not having to hire temporary employees and associated productivity savings were included in the cost evaluation.

The safety staff summarized the evaluation data in a memo to the members of the committee that worked on the belt use promotion (i.e., representatives

of the unions and management). The memo reported that the belt use promotion was a success in terms of increased belt use. The memo also included a summary of unsolicited anecdotal reports of support for the program made to the safety office by employees. Because the evaluation demonstrated the program's success, the program committee decided to conduct annual safety belt promotion programs to maintain belt use at the high levels achieved by the initial program. Without a systematic program evaluation, management and safety staff would have been unable to determine the effects of the program, and thus would have been unable to make an informed decision on whether or not to continue similar programs in the future.

The second example is a program to promote bicycle helmet use in elementary schools. The local elementary school district decided to promote bike helmet use as part of an overall injury control effort. The helmet use promotion program consisted of a 1-week bike safety education and training program conducted in 3rd grade physical education classes. The goals of the program were to increase knowledge of bike safety and the need for helmet use, to increase the ownership of helmets, and to increase use of bike helmets.



To evaluate their program, the district selected a *pre-post with control group* design. The pre-post component was deemed to be important so that program effects could be measured from a baseline condition. A control group condition was selected because the district felt it was important to control for possible effects other than the program itself which may affect students' bike safety and helmet use knowledge or behavior during the program. The district decided against having different classes within a single school serve as either a program or control group because they felt there was a strong chance that students within the same school would talk about the program with each other, and thus the control group would be "contaminated" with the program.

The two schools that were selected to pilot the program were the most separate geographically. This was done so that interactions between students of the two schools would be minimized, thus reducing the chance that the program at one school would affect knowledge or behaviors at the other. Because of the serious implications of bike helmet nonuse (i.e., serious head injury), the district decided they could not ethically withhold the program from the

control group. They decided they would implement the bike safety program in the "control" school in the semester after the first school ended its program. That is, the first school would receive the program in the fall semester and the second school would receive the program in the spring semester.

The district decided against a *pre-post only* design because they wanted to be sure that it was the program that caused any observed effects rather than some other influence the kids may be exposed to. A simple pre-post design would not permit the district to determine whether or not effects from outside their program influenced helmet use knowledge and behavior. The district also decided against using a *reversal* or *multiple baseline* design because the educational program was not expected to have a measurable effect until it was nearly completed. Thus, it would not be possible to see program effects by repeated observations throughout the program period. In addition, the district did not want to take up valuable class time by repeatedly testing the kids' knowledge of bike safety and helmet use, or constantly query them about whether or not they had purchased a helmet or when they used a bike helmet. In fact, one teacher pointed out that repeated measurement for such information was an intervention by itself, and the constant measurement could mask or exaggerate possible effects of the education program.

A single survey instrument was designed for implementation in the classes at both schools. Teachers determined it would be too difficult and costly to try to actually observe helmet use among 3rd grade students, so self-report measures of helmet use were used. The survey served two purposes. First, the survey asked questions that provided information about what students knew and did not know about bike safety and helmet use. Second, the survey asked students if they owned a helmet, had a helmet available for use, and asked about helmet use patterns. Each of the items in the survey were multiple choice, no open ended questions were used.

Students in each school were surveyed at three concurrent points in time. Students at both schools were surveyed in the fall immediately before the program began at the first school and one week after the conclusion of the program at the first school. Students at both schools were surveyed again one week after the conclusion of the program at the second school in the spring. Note that in this evaluation, the second school (the one with the program in the spring) serves as the control group for evaluating the effects of the program conducted at the first school. The second school really has no control group on which to base a detailed evaluation of the program conducted there. An evaluation of the program at the second school is a simple *pre-post* design (if this school is considered by itself). However, recall that the evaluation is

Table III
Effects of Bike Safety Education Program

		Pre-Program	Fall Post-Program	Spring Post-Program
1 	Knowledge	3.2	4.6	4.0
	Helmet ownership	2.1	3.7	3.8
	Helmet use	2.0	4.0	4.1
2 	Knowledge	3.3	3.2	4.7
	Helmet ownership	2.0	2.1	3.9
	Helmet use	1.8	1.9	4.1

not of the program as it is implemented at each school, but rather of the more general program itself.

For the sake of simplicity, let's say the knowledge, helmet ownership, and helmet use data can each be summarized on 5-point scales (1 = poor. . . 5 = excellent). In reality, each measure would probably have its own unique scale which would be more meaningful. The more simple 5-point scale is used here for illustrative purposes. Data from the schools might resemble the pattern shown in Table III. The figures given in Table III are average scale values for each school on each of the three scales (knowledge, helmet ownership, and helmet use). While the variance of scores around these means is not provided, the variance is quite important for estimating the statistical significance of differences between the means.

The school district was excited that they found increases in knowledge, reported helmet ownership, and helmet use from preprogram levels, but members felt it was important to assess if the differences were statistically significant or simply the result of chance. Before beginning the program, the district had identified a teacher who had a background in simple statistics. Working with this teacher, the district collected data which could be analyzed using a statistical procedure called ANalysis Of VAriance (ANOVA). ANOVA determines whether differences in means (averages) are statistically different or could be attributed to chance differences expected within the variance around the means.

The ANOVA analysis found that scores for School 1 increased significantly from Pre-program to Fall Post-program levels. Although there was an apparent decline from Fall Post-program to Spring Post- program in knowledge

for School 1, these differences were found to be nonsignificant (meaning that knowledge scores did not really decline). District officials were confident that it was the program which caused the increase in scores because while scores increased for School 1 (Pre- program to Fall Post-program), scores for School 2 did not change. If School 2 scores also increased during that time period, then the change in scores for School 1 could not be attributed to the bike safety program alone.

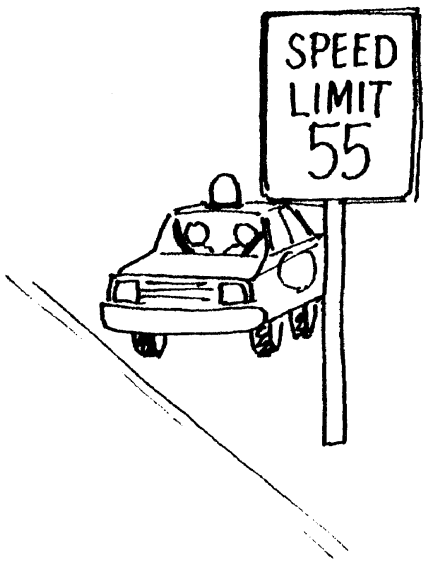
Confidence in the effectiveness of the bike safety program was strengthened by a significant increase in scores from Fall Post-program to Spring Post-program for School 2. Although there is no control group available to compare the scores from School 2 to, the fact that there was a statistically significant increase at School 2 after the program was completed strengthens the argument that the program was successful at achieving its goals of increased knowledge, helmet ownership, and helmet use. Results of the evaluation were drafted in memo form and distributed throughout the state elementary education network so that other schools could learn about the program's effectiveness and implement similar bike safety programs.

The third example is an enforcement program designed to reduce travel speeds and increase safety belt use on a stretch of highway in Apple County. The goals of the program were to:

(1) reduce average travel speeds on the highway to legal levels (i.e., 55-mph),

(2) increase safety belt use of motorists on the highway, and

(3) reduce the number and severity of crashes and crash-related injuries on the highway.



These goals were to be achieved by setting up enforcement patrols dedicated to speed limit enforcement. Once drivers had been stopped for speeding violations, all safety belt nonusers were to be issued a second violation for belt nonuse. In addition, roadside signs were posted informing drivers of the increased patrols, local broadcast media repeated public service announcements describing the program, and press releases describing the number of citations issued and the short-term effects of the program on belt use and speeds were issued each month.

Table IV
Effects of Speed and Belt Use Enforcement Program

	Speed	Belt use	Crash freq.	Injury freq.	Traffic volume	Crash rate	Injury rate
May	63.2	49%	50	15	10K	.0050	0.30
June	59.3	53%	45	12	12K	.0038	0.27
July	56.2	64%	43	10	13K	.0033	0.23
August	56.3	60%	42	10	13K	.0032	0.24

The program was scheduled to go into full force for the summer travel season (June through August).

The police decided to use a simple *pre-post* design to evaluate the program. They would have liked to have included a control group to the design, but it was found to be prohibitively costly to gather speed and belt use data on a similar highway in a second county. A reversal design was ruled out because the reversal would occur in the fall. Driving habits and crash patterns differ between the fall and summer months, thus little would be learned from a fall reversal period. On the other hand, great attention was paid to recording administrative data associated with the enforcement project. Police recorded the number of hours of dedicated patrol for the project, the number of speeding and belt nonuse violations issued, as well as the number of other citation types that were issued by officers during their patrols.

The police enlisted the assistance of the local road commission for collecting data on speeds and traffic volumes. Fortunately, the road commission had traffic count and speed data for the highway for three years prior to the enforcement program start-up. The road commission agreed to provide those data as well as to begin collecting speed data more frequently beginning one month prior to the enforcement program start-up (May through August). Travel speed and traffic counts were monitored using pneumatic tubes which crossed the road. Radar units were not used to measure speeds for the evaluation component (although they were used by police in enforcement) because the use of radar detectors tends to slow traffic in areas where radar speed monitoring devices are used. Data on crash involvement and injury severity were collected and regularly summarized by the police on an on-going basis for five years prior to program start-up. These activities continued throughout the program and beyond. Safety belt use data were gathered through bimonthly observations (May through August) conducted at high-

way intersections and off-ramps by local Explorer Scouts who were affiliated with the police and student Police Cadets.

Data collection activities went off without a hitch. While program supervisors monitored the data regularly throughout the program to find particular strengths and weaknesses, at the end of August the police wanted to review the effects of the program over the entire program period. Table IV presents results from the program. (Note that these figures are fictitious and should not be used as a basis of comparison for a real program; the results in the table are for illustrative purposes only.) Because of the low number of fatalities and serious injuries that typically occur on relatively short stretches of highway, analysis of injury frequencies and injury rates per crash included all injuries (fatal through minor).

The data show that the enforcement program was successful in achieving its goals. Average speeds declined from 62.3 mph in May prior to the program to 56.3 mph in August the final month of the program. Safety belt use increased from 49% in May to a high of 64% in July before dropping off slightly to 60% in August. While the total *number* of crashes and crash injuries declined during the program, the officer in charge of the program evaluation knew that it was more appropriate to examine the rate of crashes per the number of vehicles travelling on the roadway and the rate of injuries per vehicle crash. Both of these rates show declines from the preprogram levels in May to the levels at the end of the program in August.

The evaluation also examined the number of officer hours spent on the dedicated speed/belt use patrols, the number of speeding, belt nonuse, and other violations written, and costs associated with the patrols. This administrative evaluation found that while the program was costly in terms of officer hours and overtime, the program was effective in achieving its goals.

A report summarizing these findings was written and delivered to the police Chief. He was so encouraged by the results he contacted other local police, sheriff, and state police divisions which had this highway as part of their jurisdiction. These agencies applied for a cooperative grant from the state highway safety office for funding to support a joint enforcement effort along the entire highway. Because effectiveness data were available and demonstrated the viability of such a program, the grant was approved and the program was expanded to cover the entire highway.

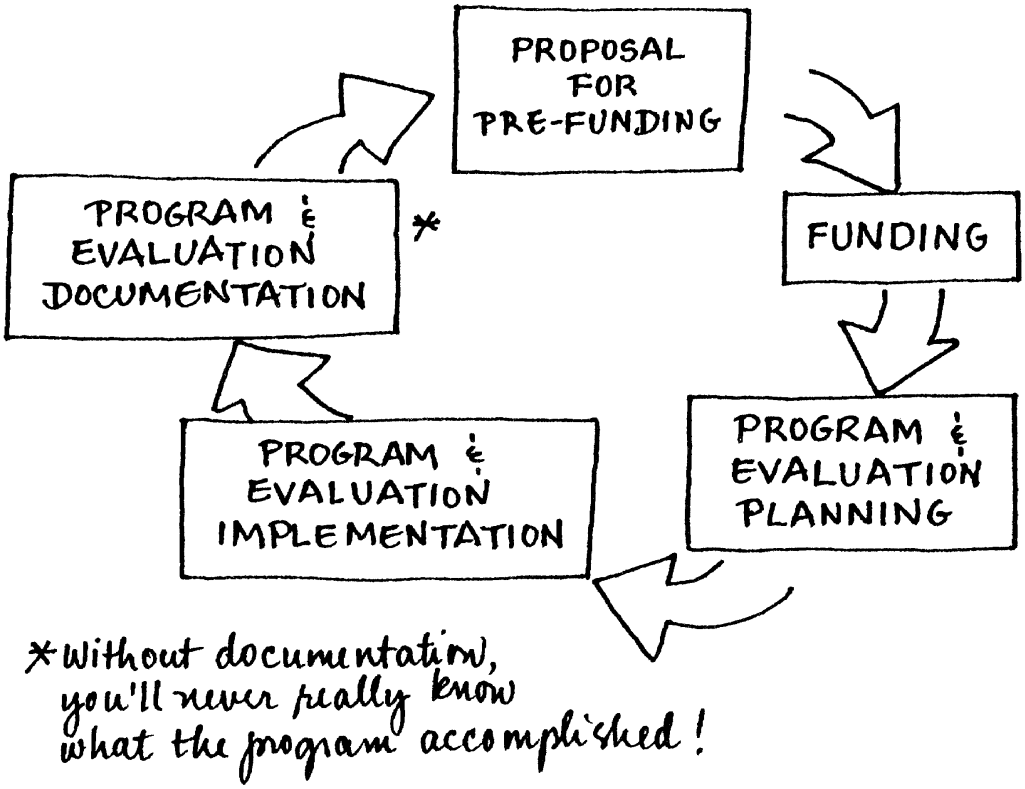
CHAPTER 8
Presenting Evaluation Results

The Confederation for Appropriate Child Seat Use distributed pamphlets describing proper installation and use of child safety seats as a project sponsored by the state highway safety office. These pamphlets were delivered to pediatricians and health clinics where children age 0-4 years are commonly treated. The confederation conducted child restraint observations at the doctors' offices and at local shopping centers frequented by parents with young children. A set of charts showing child restraint use prior to and following the program was sent to the sponsor with a one page memo thanking the sponsor for its support. Because important details about the evaluation design and how the actual program was conducted were not included in the report to the sponsor, the project was not funded in the next fiscal year. The rejection letter to the confederation said that there was insufficient evidence of program effectiveness from the prior year's program to warrant program renewal.

As important as evaluation is, it is equally important that results from evaluations be presented to sponsors, program participants, and the media in a way that meets the needs of the user. We must keep in mind that different users have different needs and uses for evaluation results. The presentation of results must be linked to the purpose for the evaluation, and be able to answer key questions raised by the program evaluation.

WRITING FOR THE SPONSOR

Sponsors are interested in evaluation results because they want to know what they got for their money, whether or not to continue to fund your project, and whether or not to expand funding for similar projects. Typically sponsors require the most detailed summaries of evaluation efforts. It is also generally in your best interest to fully detail project events and evaluation results. If



something worked, you will want to make sure the sponsor knows it worked, and their money was well spent. If something did not work as planned, you will want to detail what may have prevented its full success so future funding and program efforts are not hampered by similar problems.

A detailed evaluation report has the following sections: (1) executive summary, (2) background and introduction, (3) methods, (4) results, and (5) discussion. The executive summary should provide a brief overview of the evaluation, providing background for the project and evaluation, listing its major findings, and summarizing conclusions of the evaluation. Because many people will read only the executive summary, it should be sufficiently detailed that readers can understand the project and its evaluation. The executive summary is most often read by people too busy or hurried to read the entire report; therefore, it should be no longer than 2-3 pages at the most. Although the executive summary is generally placed first in the report, it is generally written last to ensure it accurately reflects the positions taken in the full report.

BACKGROUND & INTRODUCTION. The background and introduction describes why you conducted the program and details the specifics of the program itself. It should begin with a brief discussion of what prompted the program (e.g., high number of alcohol-related crash deaths, low safety belt use). Once you have laid the groundwork for why the program was conducted

you should describe the program itself. Begin by listing the specific goals and objectives of the project. Describe what the program was designed to accomplish.

Once you have described the background and goals of the program, detail the specifics of the program itself. Describe the target groups the program was designed to affect. Detail the procedures or protocols used to implement the program. Include copies of materials used for the program (handouts, brochures, posters, news clippings, etc.) Describe the resources that were used to carry out the program (e.g., program sponsors, prize donors, technical assistants, volunteers). Describe administrative procedures that were required and obstacles that had to be overcome to conduct the program. Detail specific times, dates, and places of key events in the program. A rule of thumb for the most thorough types of reports: *you have provided enough detail about your project if you can reproduce the project exactly from the detailed description in the report.* While a report with this much detail often seems to be excessive, this level of detail will be useful for you by documenting your program for future reference.

METHODS. The methods section should describe the methods used in the evaluation of the project. It should begin with a description of the goals of the evaluation. It is useful to detail the specific questions the evaluation was intended to answer. The rest of the section details how the program was evaluated. You should describe the evaluation design and why that design was selected (e.g., explain why a reversal design was chosen instead of a control group design). You should also detail the strengths and weaknesses associated with your design selection. That is, what can and cannot be said about the results you will get using the design selected.

You must detail the data collection procedures used. Describe the sampling procedure that was used to select people, groups, or sites for measurement. Detail how data were collected (time, place, method of data collection). If you use an automated data collection device (e.g., breath alcohol sensor, speed monitoring devices, etc.) describe the machine in detail so readers understand why it was used and what information it provided. Also list the make and model of the device along with the name of the manufacturer so readers know exactly what device was used. Include copies of data collection forms and surveys in the report. Include a copy of the survey instrument, observational data collection form, and any other data collection instrument in the appendix at the end of the report. Detail training given to persons who collected the data or operated the data collection devices.

In the methods section you should also detail the data analysis techniques used

to examine the data. If you only charted the data and used a visual inspection procedure to determine program effects, say so. If you used a more complex analytic plan, describe what your analysis consisted of using terms that the readers of the report will understand. Remember that much of your audience will be unfamiliar with many analytic techniques. Simplify your discussion of analytic techniques so that the uninitiated can understand what you did. Brief examples are helpful for explaining complex points.

RESULTS. Details of the actual observations, measurements, other data collection procedures, and analyses of these data are described in the results section. In this section you should include charts, graphs, and tables which describe the data. Present the results in such a way that readers know the answers to the evaluation questions detailed in the beginning of the methods section. Results should be kept as straightforward as possible. Detailed discussion of possible ramifications of the results is best left for the discussion section. However, you should include brief interpretations of the results in this section. These brief interpretations help readers understand the importance and meaning of the results. For example, *we found a statistically significant increase in test scores between time 1 and time 3 indicating the program was effective in educating drivers about the need to wear safety belts even in vehicles equipped with airbags.*

DISCUSSION. The discussion section is your opportunity to detail your interpretation of the results and their ramifications. This is a departure from the results section in which discussions should remain brief and to the specific point raised by the result being considered. In the discussion section you should bring together all of the results to generate more global interpretations of the evaluation. Discuss your conclusions about the program based on the program goals and findings of the evaluation. Describe what you believe went right and what went wrong with both the program and evaluation. Speculate on how improvements could be made for the future. Describe what the next step in evaluation or programming should be to achieve those goals which were not met. If all your goals were met, describe the next steps for program dissemination or expansion. Give the sponsor your suggestions for the future based on your evaluation.

WRITING FOR PROGRAM PARTICIPANTS

Program participants and sponsors both typically want answers to the same questions. However, participants often prefer more simple, straightforward answers to the key questions addressed in the evaluation. When writing for program participants you should consider

simplifying the evaluation report to address the specifics of the program and its outcome.

The report for participants should be structured in the same basic format as the report to the sponsor, but the participant's report may be written in a more direct style. You generally can't go wrong by writing a detailed report for both audiences, but often participants and their decision makers are more interested in just the "bottom line" than sponsors. That is, they want to know if the program worked, and if not, how to fix future programs to increase the likelihood that they will work. Often program participants are more sophisticated about the specifics of the program they are involved with and less sophisticated about evaluation techniques than are program sponsors (of course the opposite is also true on occasion). Reports should always consider the level of sophistication of the audience.

WRITING FOR THE MEDIA

The media can be your ally or your foe. The key to which side they seem to be on is generally determined by how you present your program results. Media reports are almost always constrained by time or space. Reports to the media should always be concise and to the point. The likelihood that a story about your program will be covered by the broadcast or print media is often governed by how simply you can convey your results and how relevant the material can be made to seem to an audience.

When presenting material for television, remember that TV is a visual medium. Try to present the material in a way that easily lends itself to moving pictures. Television time is extremely valuable and even important stories usually get only brief coverage. Make sure you present your work in such a way that the media can describe the program and its results quickly. The same goes for radio and newspaper articles as well.

The media prefers simple yes-no, good-bad answers to questions more than answers that require endless qualifications for accuracy. Even if you don't give the media a yes-no answer, they will often generate one from your comments. The key is to anticipate questions and develop brief, simple answers that can be used without significant revision. This often means oversimplifying your results. Remember, if you don't do it, often they will. Another hint, despite what you say, treat everything as though it is "on the record." While most reporters are honorable, "off the record" comments have a way of sneaking into reports.

The media also likes "catchy" phrases. When writing a press release or giving an interview have a few "quotable phrases" ready to throw out. Often these are points that emphasize the dramatic or are written or spoken using dramatic language. You can never really predict what phrases will catch a reporter's ear, but they tend to be simple, brief, and conclusive (that is, not wordy, overanalytic responses).

CHAPTER 9

Identifying Resources for Evaluations

The Apple County Traffic Safety Committee wanted to evaluate a safety belt promotion campaign planned for the county. The program consisted of several separate projects promoting belt use in the community. The committee was interested in evaluating effects of the individual projects in addition to the overall effect of the larger program in the county. However, the committee believed it did not have sufficient financial resources to sponsor both the projects and their evaluations. The decision was made that projects had a higher priority than evaluation. Thus, little effort was made to evaluate either the individual projects or the overall effect of the total program. Because there was no evaluation, the committee could not make an informed decision of whether or not to spend funds in the coming fiscal year for similar projects. The decision was made not to continue because the committee decided that the belt use rate had to be high because they had spent so much last year to increase it.

Often projects go without evaluation because of the belief that there are insufficient resources available to conduct an evaluation. Conducting project evaluations need not be expensive or complex. Often conducting an effective evaluation is simply a matter of identifying resources in the community available to provide assistance. Even if sufficient resources for a complete and thorough "textbook" evaluation are unavailable, project staff should always conduct evaluations as best they can with available resources.

One obvious source of evaluation resources is the agencies involved with the program. It is always a good idea to check with each of the involved agencies and maximize use of resources available from each agency. In fact, it is a good idea to try to foresee evaluation needs for your project and involve agencies who may be of assistance from the very beginning.

Most agencies involved with transportation (e.g., police, road commissions, departments of transportation, driver licensing agencies, courts) are as concerned about safety issues as your organization. Involve them in your program from the start. Many regularly collect data that may prove useful to your evaluation. Some have personnel well versed in techniques of program evaluation and data analysis who can be of assistance.

An excellent resource for information on traffic safety issues is the highway safety office or bureau in each state.

An excellent resource for information on traffic safety issues is the highway safety office or bureau in each state. These agencies often are able to bring together groups with common goals who can work together to enhance program effectiveness. These highway safety agencies also typically know about available resources to assist you with data collection and analysis. In Michigan, the office is called the Office of Highway Safety Planning; it is the Governor's Highway Safety Office. The name and affiliation for these offices differs from state to state, but they all can be quite helpful to you and your projects.

The most costly types of evaluations are generally those that involve a lot of direct observation of behavior. Behavioral observations are often time consuming. Some require that many observers be in place making observations at the same time (e.g., one-shot safety belt observations over a large geographic area). There are a variety of service groups which can be of assistance with data collection efforts. Contact the scouts, service organizations like Rotary or the Jaycees, senior citizens groups, or high school service clubs and discuss your program with them. They are often receptive to assisting with traffic safety programs. Don't be discouraged if you are turned down by some of these organizations. Each has a mission independent of your project, and your program might not fit that mission or their available time table.

When contacting volunteers for assistance, remember that even volunteers require something in exchange for their service. Often all they require is a pat on the back and a "job well done." Other times they will require that their organization be recognized for their efforts in press releases or reports of the project. This is usually a good exchange. You get some help, they get some press. On occasion, more substantive incentives are needed to encourage participation. Try to solicit donations from local merchants to reward the volunteers (e.g., fast food coupons). Many local businesses are happy to

donate small incentives to promote their business. Sometimes you may find the organization will perform the desired service for a donation to their organization. Don't turn them down immediately. Check your budget and determine what you can afford. When a donation is requested it usually amounts to substantially less than you would have to pay if you hired people off the street or paid your staff to do the job.

When contacting volunteers for assistance, remember that even volunteers require something in exchange for their service.

The skill most organizations lack is expertise in data analysis. This is particularly true for analyses requiring application of statistics. Don't immediately dismiss evaluation strategies that require statistical analyses if no one in your organization has statistical expertise. Often you can find individuals who can help you in local colleges, universities, community colleges, or even advanced high school classes. Occasionally local agencies involved with traffic safety have persons who can assist you in your data analyses. Give these folks a call. College professors and teachers are often eager to become involved in projects which illustrate applications of techniques to students.

CLOSING COMMENTS

This manual was intended to help people to better design, implement, and report evaluations of their traffic safety programs. We discussed many of the most important topics in evaluation research here. However, there are still many subject areas and subtleties of research that were not discussed. Use this manual as it was intended, as a guide to the basics of evaluation. We would have liked to have provided you with a "cookbook" scheme for evaluating traffic safety programs: if you have program type A and B, use design #1 with data collection plan #2." Unfortunately, there are too many idiosyncratic features in every program to make such a cookbook possible. However, we have provided general "rules of thumb" you should find helpful in your evaluation efforts.

Traffic safety can be a valuable and rewarding field of endeavor. Its rewards are magnified when you have data to support your successes and have the information necessary to overcome your disappointments.