Demonstrating Your Program's Worth

A Primer on Evaluation for Programs to Prevent Unintentional Injury

Nancy J. Thompson, PhD
Helen O. McClintock

National Center for Injury Prevention and Control
Atlanta, Georgia
1998
Second Printing (with revisions), March 2000

Demonstrating Your Program's Worth is a publication of the National Center for Injury Prevention and Control, Centers for Disease Control and Prevention:

Centers for Disease Control and Prevention
Jeffrey P. Koplan, MD, MPH, Director

National Center for Injury Prevention and Control
Stephen B. Thacker, MD, MSc, Director

Division of Unintentional Injury Prevention
Christine M. Branche, PhD, Director

Production services were provided by the staff of the Office of Health Communication Resources, National Center for Injury Prevention and Control:

Graphic Design
Marilyn L. Kirk

Cover Design
Beverly Charday
Mary Ann Braun

Text Layout and Design
Sandra S. Emrich

Suggested Citation: Thompson NJ, McClintock HO. Demonstrating Your Program's Worth: A Primer on Evaluation for Programs To Prevent Unintentional Injury. Atlanta: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, 1998.

Purpose

We wrote this book to show program managers how to demonstrate the value of their work to the public, to their peers, to funding agencies, and to the people they serve. In other words, we're talking about how to evaluate programs—a scary proposition for some managers. Our purpose is to reduce the scare factor and to show that managers and staff need not be apprehensive about what evaluation will cost or what it will show.

Remember that there are two ways an injury-prevention program can be successful. The obvious way is if it reduces injuries and injury-related deaths. The other way is if it shows that a particular intervention does *not* work. A program is truly worthwhile if it implements a promising intervention and, through evaluation, shows that the intervention does *not* reduce injuries. Such a result would be of great value to the injury prevention community: it would save you and other programs from wasting further resources and time on that particular intervention.

In this book, we show why evaluation is worth the resources and effort involved. We also show how to conduct simple evaluation, how to hire and supervise consultants for complex evaluation, and how to incorporate evaluation activities into the activities of the injury prevention program itself. By learning to merge evaluation and program activities, managers will find that evaluation does not take as much time, effort, or money as they expected.

Acknowledgments

We acknowledge and appreciate the contributions of several colleagues: Dr. Suzanne Smith saw the need for this primer and began the project; Drs. Terry Chorba and David Sleet enumerated the various types of injury programs (page 73); and Dr. Jeffrey Sacks, Dr. Katherine Miner, Dr. David Sleet, Ms. Susan Hardman, and Mr. Roger Trent reviewed the content and provided invaluable suggestions.

Contents

Introduction

All too often public health programs do wonderful work that is not properly recognized by the public, by other health care professionals, or even by the people who benefit directly from the program's accomplishments. Why should this be? In most cases, it is because program managers and staff strongly believe that their work is producing the desired results but have no solid evidence to *demonstrate* their success to people outside their program. In other words, such programs are missing one key component: evaluation.

Unfortunately, without objective evaluation, program managers and staff cannot show that their work is having a beneficial effect, and other public health programs cannot learn from their success.

In addition, without adequate evaluation, programs cannot publish the results of their work in medical, scientific, or public health journals, and they cannot show funding agencies that their work is successful. Obviously programs that produce facts and figures to prove their success are more likely to publish the results of their work and more likely to receive continued funding than are programs that cannot produce such proof.

And here is another important point about evaluation. It should begin while the program is under development, *not* after the program is complete. Indeed, evaluation is an ongoing process that begins as soon as someone has the idea for a program; it continues throughout the life of the program; and it ends with a final assessment of how well the program met its goals.

Why must evaluation begin so early? Consider, for example, if you were to set up a program to provide free smoke detectors to low socioeconomic households. You put flyers in the mail boxes of people you want to reach, inviting them to come by your location for a free detector. Many people respond but not as many as you expected. Why?

To find out, you evaluate. Perhaps you learn that your location is not on a bus line and many people in your target population do not own cars. Or, perhaps, the language in

your flyer is too complex to be easily understood by the people you want to read it. So you rewrite your flyer or move your location. Would it have been better to test the language in the flyer for readability and to assess the convenience of your location *before* beginning the program? Yes. It would have saved time and money—not to mention frustration for the program staff.

So, the moral is this: evaluate, and evaluate *early*. The earlier evaluation begins, the fewer mistakes are made; the fewer mistakes made, the greater the likelihood of success. In fact, for an injury prevention program to truly show success, evaluation must be an integral part of its design and operation: evaluation activities must interweave with—and sometimes merge into—program activities. If a program is well designed and well run, evaluating the final results can be a straightforward task of analyzing information gathered while the program was in operation. In all likelihood, the results of such an analysis will be extremely useful, not only to your own program but to researchers and to other injury prevention programs.

To help program managers avoid difficulty with evaluation, we produced this primer. Its purpose is to help injury prevention programs understand 1) why evaluation is worth the resources and effort involved, 2) how evaluation is conducted, and 3) how to incorporate evaluation into programs to prevent unintentional injuries. This primer can also help program managers conduct simple evaluation, guide them in how to hire consultants for more complex evaluation, and allow them to oversee the work of those consultants in an informed way.

Our purpose is to help injury prevention programs understand 1) why evaluation is worth the resources and effort involved, 2) how evaluation is conducted, and 3) how to incorporate evaluation into programs to prevent unintentional injuries.

Since we want to practice what we preach, we ask that you help us with our evaluation of this book. We encourage you to give us your opinion. Is this book useful? If so, how have you found it useful? Are all sections clear? If not, which sections are unclear? Is the book's organization easy to follow? If not, where have you had difficulty? Should we add more details? If so, on which topics? We are interested in any comments or suggestions you might have to improve this book and make it more useful. Your close involvement with the people that you and CDC want to serve makes your feedback invaluable.

On page 125 is a form you can use to send us your comments. We look forward to hearing from you.

Please also visit our web site for more information about injury control and prevention and to order a variety of free publications: www.cdc.gov/ncipc/pub-res/pubs.htm

How This Primer Is Organized

This book is designed to help program staff understand the processes involved in planning, designing, and implementing evaluation of programs to prevent unintentional injuries.

Section 1 has general background information explaining why evaluation is important, what components go into good evaluation, who should conduct evaluation, and what type of information evaluation will provide.

"Evaluation at a Glance" (page 23) is a quick reference that helps programs decide when to conduct each stage of evaluation, describes what kind of information each stage of evaluation will produce, and explains why such information is useful.

In **Section 2**, we describe each of the four stages of evaluation: formative, process, impact, and outcome. In particular, we discuss the appropriate time to conduct each stage and the most suitable methods to use. "Evaluation at a Glance" (page 23) is a quick reference that helps programs decide when to conduct each stage of evaluation, describes what kind of information each stage of evaluation will produce, and explains why such information is useful. For further help in deciding which stage of evaluation is appropriate for your program, we guide you through a set of questions (page 24). Your answers will tell you which stage is the right one for your program's situation.

Section 3 is devoted to the methods for conducting evaluation. We provide enough information to enable you to conduct simple evaluation. However, the primary use is to enable you to communicate with, hire, and supervise evaluation consultants.

Appendix A contains sample questions for interviews, focus groups, and questionnaires. It also contains sample events to observe and items to count at certain stages of evaluation. These examples can be adapted for use in evaluating any program to prevent unintentional injury.

Appendix B contains sample forms to help keep track of contacts that the program makes with the target population, items received from the target population, and items dispensed during a product distribution program.

Appendix C is a checklist of tasks that all programs to prevent unintentional injury can follow to make sure they do not omit any evaluation step during program design, development, and implementation.

Appendix D contains a bibliography of sources for further information about various aspects of evaluation.

Appendix E is a glossary of terms used in this primer.

On page 125 is a form that you can use to send us comments about this book.

SECTION 1

GENERAL INFORMATION

General Information

Introduction

Evaluation is the process of determining whether programs—or certain aspects of programs—are appropriate, adequate, effective, and efficient and, if not, how to make them so. In addition, evaluation shows if programs have unexpected benefits or create unexpected problems.[1]

All programs to prevent unintentional injury need to be evaluated whether their purpose is to prevent a problem from occurring, to limit the severity of a problem, or to provide a service.

And evaluation is much easier than most people believe. A well-designed and well-run injury prevention program produces most of the information needed to appraise its effects. As with most tasks, the key to success is in the preparation. Your program's accomplishments—and the ease with which you can evaluate those accomplishments—depend directly on the effort you put into the program's design and operation.

Ah, there's the rub: whether 'tis wiser to spend all your resources running the injury prevention program or to spend some resources determining if the program is even worth running. We recommend the second option: programs that can demonstrate, through evaluation, a high probability of success also have a high probability of garnering legislative, community, technical, and financial support.

History of Evaluation

Early attempts to evaluate programs took one of two forms:

- Evaluation based on practical experience.
- Evaluation based on academic rigor.

Practical evaluation was conducted by people involved in prevention programs or by program staff. They were careful not to disrupt program activities more than absolutely necessary and to divert as few resources as possible away from the people being served. As a result, the evaluation design was often weak, and the data produced by the evaluation lacked credibility.

In contrast, academic evaluation was, in general, well designed and rigorously conducted. However, it was labor-intensive, intrusive, and therefore not applicable to large portions of the population because the results represented only the knowledge, attitudes, beliefs, or behaviors of people who would complete a laborious regimen of evaluation procedures.[2]

Over the years, evaluation evolved. The discipline profited from both practical experience and academic discipline. Methods became more feasible for use in the program setting and, at the same time, retained much of their scientific value. Furthermore, we now understand that effective evaluation begins when the idea for a program is conceived. In fact, much of the work involved in evaluation is done while the program is being developed. Once the prevention program is in operation, evaluation activities interact—and often merge—with program activities.

Much of the work involved in evaluation is done while the program is being developed. Once the prevention program is in operation, evaluation activities interact—and often merge—with program activities.

Purpose of Evaluation

Data gathered during evaluation enable managers to create the best possible programs, to learn from mistakes, to make modifications as needed, to monitor progress toward the program's goal, and to judge the program's ultimate outcome (Figure 1).

Indeed, *not* evaluating an injury prevention program is irresponsible because, without evaluation, we cannot tell if the program benefits or harms the people we are trying to help. Just as we would not use a vaccine that was untested, we should not use injury interventions that are untested. Ineffective or insensitive programs can build public resentment and cause people to resist future, more effective, interventions.

Evaluation will also show whether interventions other than those planned by the program would be more effective. For example, program staff might ask police officers to talk to

students about the hazards of drinking and driving. The hope might be that stories the police tell about the permanently injured and dead teenagers they see in car crashes would scare the students into behaving responsibly.

Evaluation might show, however, that many teenagers do not respect or trust police officers and therefore do not heed what they say. Evaluation would also show what type of people the students would listen to—perhaps sports stars or other young people (their peers) who are permanently injured because of drinking and driving.

The right message delivered by the wrong person can be nonproductive and even counterproductive.

Why Evaluate Injury-Prevention Programs?

- To learn whether proposed program materials are suitable for the people who are to receive them.
- To learn whether program plans are feasible before they are put into effect.
- To have an early warning system for problems that could become serious if unattended.
- To monitor whether programs are producing the desired results.
- To learn whether programs have any unexpected benefits or problems.
- To enable managers to improve service.
- To monitor progress toward the program's goals.
- To produce data on which to base future programs.
- To demonstrate the effectiveness of the program to the target population, to the public, to others who want to conduct similar programs, and to those who fund the program.

Figure 1.

Side Benefits of Evaluation

A side benefit of formal evaluation is that the people who are served by the program get an opportunity to say what they think and to share their experiences. Evaluation is one way of listening to the people you are trying to help. It lets them know that their input is valuable and that the program is not being imposed on them.

Evaluation is one way of listening to the people you are trying to help.

Another side benefit is that evaluation can boost employee morale—program personnel have the pleasure of seeing that their efforts are not wasted. Evaluation produces evidence to show either that their work is paying off or that management is taking steps to see that needed improvements are made.

A third side benefit is that, with good evaluation before, during, and after your program, the results may prove so valuable that the news media or scientific journals will be interested in publishing them. In addition, other agencies or groups may see how well you have done and want to copy your program.

A Common Fear of Evaluation: *"It shows only what's wrong!"*

Often, a major obstacle to overcome is the program personnel's concern that evaluation will reveal something bad that they are unaware of. And the truth is that evaluation will reveal new information that shows any aspects of the program that are not working as effectively as planned. But that is not bad news; that is good news. Because now something can be done to improve matters. We promise that evaluation will also bring news about aspects that are working better than expected.

Indeed, all evaluation produces four categories of information (See Table 1).

Table 1. Four Categories of Information Produced by Evaluation

	Characteristics	Description
1.	➤ Information staff knows already. ➤ Indicates program is working well.	Data about aspects of program that work well, that program staff knows about, and that should be publicized whenever possible.
2.	➤ Information staff knows already. ➤ Indicates program needs improvement.	Data about aspects of program that need improvement, that staff knows about and hopes will not be found out. Staff is unlikely to mention these aspects to the evaluator.
3.	➤ New information. ➤ Indicates program is working well.	Data about aspects of program that work well, but staff does *not* know about them. All evaluation uncovers some pleasant surprises, but program staff rarely expects them.
4.	➤ New information. ➤ Indicates program needs improvement.	Data about aspects of program that need improvement and about which staff is unaware. This is the type of information staff most expects when evaluation begins.

All evaluation uncovers some pleasant surprises, but program staff rarely expects them.

Well-designed evaluation always produces unexpected information. That information is just as likely to be about something that works well as it is to be about something that needs improvement.

So remember to expect pleasant surprises; and recognize that, by showing you *why* certain components of your program do not work, evaluation will often make what seemed an intractable problem easy to solve. With this change in perspective, evaluation ceases to be a threat and becomes an opportunity.

Choosing the Evaluator

The first step in any evaluation is deciding who will do it. Should it be the program staff or should outside consultants be hired?

In almost all cases, outside consultants are best because they will look at your program from a new perspective and thereby provide you with fresh insights. However, outside consultants do not necessarily have to come from outside your organization. Evaluators within your organization who are

not associated with your program and who have no personal interest in the results of an evaluation may serve your needs. Figure 2 contains a list of the most important characteristics of a consultant. Although these characteristics are listed with some regard to order of importance, the actual order depends on your program's needs and the objectives for the evaluation.

Important factors to consider when selecting consultants are their professional training and experience. Some specialize in quantitative methods, others qualitative. Some have experience with one stage of evaluation, others with another stage. Some consider themselves in partnership with program staff; others see themselves as neutral observers. Some had formal courses in evaluation; others learned evaluation on the job. In other words, the background experiences of evaluators can vary considerably. They can even come from different professional disciplines (e.g., psychology, mathematics, or medicine). Find a consultant whose tendencies, background, and training best fit your program's evaluation goals.

Another factor to consider is the consultant's motivation (beyond receiving a fee). Consultants' personal motivations will affect their perspective as they plan and implement the evaluation. For example, some consultants may be interested in publishing the results of your evaluation and consequently *may* shade results toward what they believe would interest journal editors. Other consultants may be interested in using the findings from your evaluation in their own research (e.g., they may be researching why certain people behave a certain way). Find consultants whose professional interests match the purpose of your evaluation. For example, if the purpose of your evaluation is to ensure that the program's written materials are at the correct reading level for the people you are trying to reach, find a consultant whose interest is in producing data on which management can base decisions.

Listed next are some areas consultants specialize in:

- Conducting basic research.
- Producing data on which managers can base decisions (data may cover broad social issues or focus on a specific problem).
- Solving problems associated with program management.
- Increasing a program's visibility to one or more audiences.
- Documenting the final results of programs.

Make sure the consultants you hire have experience in conducting the evaluation methods you need, in evaluating programs similar to yours, and in producing the type of information you seek. Be sure to check all references before you enter into a contract with any consultant.

Characteristics of a Suitable Consultant

- Is not directly involved in the development or running of the program being evaluated.
- Is impartial about evaluation results (i.e., has nothing to gain by skewing the results in one direction or another).
- Will not give in to any pressure by senior staff or program staff to produce particular findings.
- Will give staff the full findings (i.e., will not gloss over or fail to report certain findings for any reason).
- Has experience in the type of evaluation needed.
- Has experience with programs similar to yours.
- Communicates well with key personnel.
- Considers programmatic realities (e.g., a small budget) when designing the evaluation.
- Delivers reports and protocols on time.
- Relates to the program.
- Sees beyond the evaluation to other programmatic activities.
- Explains both benefits and risks of evaluation.
- Educates program personnel about conducting evaluation, thus allowing future evaluations to be done in house.
- Explains material clearly and patiently.
- Respects **all** levels of personnel.

Figure 2.

Cost of Evaluation

Cost will vary depending on the experience and education of the consultant, the type of evaluation required, and the geographic location of your program. However, a good rule is for service programs (e.g., programs to distribute smoke

detectors to qualified applicants) to budget about 10% to 15% of available funds for evaluation.

Programs with experimental or quasi-experimental designs (see page 51) are essentially research projects, so evaluation is built into the design of the program: the cost of the program includes the cost of evaluation. Operating programs with an experimental or quasi-experimental design is more expensive than operating service programs, but experimental or quasi-experimental programs will show whether the service being provided to the target population produces the intended result. Indeed such programs are likely to produce publishable information that can benefit other programs to prevent unintentional injuries.

Be sure to include the cost of evaluation in your proposals for grant funds.

Be sure to include the cost of evaluation in your proposals for grant funds.

Designing Your Program So That Evaluation Is an Integral Part

The information needed to evaluate the effects of your program will develop naturally and almost effortlessly if you put the necessary time and resources into designing a good program, pilot testing your proposed procedures and materials, and keeping meticulous records while the program is in operation.

To be the most effective, evaluation procedures and activities must be woven into the program's procedures and activities. While you are planning the program, also plan how you will judge its success.

Include the following components in the design of your program:

- A plan for pilot testing all the program's plans, procedures, activities, and materials (see "Formative Evaluation," page 25).
- A method for determining whether the program is working as it should and whether you are reaching all the people your program planned to serve (see "Process Evaluation," page 27).
- A system for gathering the data you will need to evaluate the final results of your program (see "Impact Evaluation," page 29, and "Outcome Evaluation," page 32).

COMPONENTS OF AN EVALUATION[3]

Every evaluation must contain certain basic components (Figure 3):

- **A Clear and Definite Objective:**
  Write a statement defining clearly and specifically the objective for the evaluation.

  Without such a statement, evaluators are unfocused and do not know what to measure. The statement will vary depending on the aspect of the program that is being evaluated. For example, before the program begins, you will need to test any materials you plan to distribute to program participants. In such a case, your evaluation objective might read something like this:

  > *To learn whether the people in our target population can understand our new brochure about the benefits of smoke detectors.*

  Your evaluation objective for a completed program might read like this:

  > *To measure how many deaths were prevented as a result of our program to increase helmet use among teenage bicyclists in XYZ County.*

- **A Description of the Target Population:**
  Define the target population and, if possible, the comparison (control) group. Be as specific as possible.

  The target population will vary depending on the reason for the evaluation. In Section 2 (page 19), we discuss how to select an appropriate target population at each stage of evaluation. An example definition of a target population might read like this:

  > *All children from 8 through 10 years old who own bicycles and who attend public schools in XYZ County.*

- **A Description of What Is To Be Evaluated:**
  Write down the type of information to be collected and how that information relates to your program's objectives.

  For example, if the goal of your program is to increase the use of smoke detectors among people with low incomes,

the description of the information you need during the first stage of evaluation might read like this:

For baseline information on our target population, we need to know the number and percentage of people with incomes below $_____ in the city of XYZ who now have smoke detectors in their homes.

- **Specific Methods:**
  Choose methods that are suitable for the objective of the evaluation and that will produce the type of information you are looking for.

  In Section 2 (page 19), we discuss the four stages of evaluation and mention the methods most suitable for each stage. In Section 3 (page 35), we discuss the various methods in considerable detail.

- **Instruments To Collect Data:**
  Design and test the instruments to be used to collect information.

  In Section 3 (page 35), we discuss various methods of collecting information and the most suitable instruments for each method. For example, you could collect information on people's attitude toward wearing seatbelts by doing a survey (the method) using a questionnaire (the instrument).

- **Raw Information:**
  Collect raw information from the members of the target population.

  Raw information is simply the information you collect as you run the program (e.g., the number of people who came to your location or the number of items you have distributed). Raw information is information that has not been processed or analyzed.

- **Processed Information:**
  Put raw information into a form that makes it possible to analyze.

  Usually, that means entering the information into a computer data base that permits the evaluator to do various statistical calculations.

- **Analyses:**
  Analyzing either quantitative or qualitative information requires the services of an expert in the particular evaluation method used to gather the information.

  We discuss analysis when we describe each method in detail (see Section 3, page 35).

- **Evaluation Report:**
  Write a report giving results of the analyses and the significance (if any) of the results.

  This report could be as simple as a memo explaining the results to the program manager. However, it could also be an article suitable for publication in a scientific journal or a report to a Congressional Committee. The type of report depends on the purpose of the evaluation and the significance of the results.

Steps Involved in Any Evaluation

1. Write a statement defining the objective(s) of the evaluation.
2. Define the target population.
3. Write down the type of information to be collected.
4. Choose suitable methods for collecting the information.
5. Design and test instruments appropriate to the chosen methods for collecting the information.
6. Collect raw information.
7. Process the raw information.
8. Analyze the processed information.
9. Write an evaluation report describing the evaluation's results.

Figure 3.

SECTION 2

STAGES OF EVALUATION

Figures

Stages of Evaluation

Introduction

Ideally, evaluation is an ongoing process that begins as soon as the idea for an injury prevention program is conceived, interweaves with program activities throughout the life of the program, and ends after the program is finished. Sometimes evaluation continues for years after the program ends to see if program effects are sustained over time. By evaluating each step, programs can catch and solve problems early, which not only saves time and money but makes success more likely.

Evaluation has four stages that are begun in this order: formative, process, impact, and outcome. Planning for each stage begins while an injury prevention program is being developed, and no stage is truly complete until the program is over. Below is a brief description of each stage.

Formative Evaluation: Formative evaluation is a way of making sure program plans, procedures, activities, materials, and modifications will work as planned. Begin formative evaluation as soon as the idea for a program is conceived. Conduct further formative evaluation whenever an existing program is being adapted for use with a different target population or in a new location or setting. A program's success under one set of circumstances is not a guarantee of success under other circumstances. For evaluation purposes, an adapted program is a new program. Another occasion for formative evaluation is when an operating program develops problems but the reason is unclear or the solution not obvious. For details, see "Formative Evaluation," page 25.

Process Evaluation: The purpose of process evaluation is to learn whether the program is serving the target population as planned and whether the number of people being served is more or less than expected. Begin process evaluation as soon as the program goes into operation. At this stage, you are not looking for results. You are merely trying to learn whether

you are connecting with people in your target population as planned and whether they are connecting with you. Essentially, process evaluation involves counting all contacts with the people you are trying to reach and all events related to those contacts. For details, see "Process Evaluation," page 27.

Impact Evaluation: The purpose of impact evaluation is to measure whatever changes the program creates in the target population's knowledge, attitudes, beliefs, or behaviors. Collect baseline information for impact evaluation immediately before, or just as, the program goes into operation. Gather information on changes brought about by the program as soon as program personnel have completed their first encounter with an individual or group from the target population. For example, the first encounter might be with a person who responded to a newspaper advertisement announcing the availability of low-cost smoke detectors for people with low incomes. Or your first encounter might be with a group of people in a class on how to install a child's car seat. Impact evaluation gives you intermediate results of the program (e.g., how much people's knowledge or attitudes about restraining children in car seats have changed). For details, see "Impact Evaluation," page 29.

Outcome Evaluation: For ongoing programs (e.g., a series of safety classes taught each year to all third graders in your area), conduct outcome evaluation at specified intervals (e.g., every year, every 3 years, or every 5 years). For one-time programs (e.g., distribution of a limited number of free smoke detectors to people with low incomes), conduct outcome evaluation after the program is finished. The purpose is to learn how well the program succeeded in achieving its ultimate goal (i.e., decreasing injury-related morbidity and mortality). Such decreases are difficult to measure, however, because the rates of morbidity and death due to unintentional injuries are low. Measuring changes in events that occur infrequently takes a long time (usually years) and requires a large number of study participants. However, we will show you a way to convert data on behavior change into estimates of changes in morbidity and mortality (page 64). For details on "Outcome Evaluation," see page 32.

> *A Word of Caution:* We said above that the rates of unintentional injuries are low, which may give the impression that working to prevent them is not a good use of resources. Nothing could be further from the truth. Rates of unintentional injury may be low; nevertheless, unintentional injury is the leading cause of death for young people 1 through 34 years old and the third leading cause of death for people 35 through 54 years old.[4] Working to prevent unintentional injuries is a vital public health function.

Evaluation at a Glance

Stage 1: Formative Evaluation (For details, see page 25)

When to use:

- During the development of a new program.
- When an existing program 1) is being modified, 2) has problems with no obvious solutions, or 3) is being used in a new setting, with a new population, or to target a new problem or behavior.

What it shows:

- Whether proposed messages are likely to reach, to be understood by, and to be accepted by the people you are trying to serve (e.g., shows strengths and weaknesses of proposed written materials).
- How people in the target population get information (e.g., which newspapers they read or radio stations they listen to).
- Whom the target population respects as a spokesperson (e.g., a sports celebrity or the local preacher).
- Details that program developers may have overlooked about materials, strategies, or mechanisms for distributing information (e.g., that the target population has difficulty reaching the location where training classes are held).

Why it is useful:

- Allows programs to make revisions before the full effort begins.
- Maximizes the likelihood that the program will succeed.

Stage 2: Process Evaluation (For details, see page 27)

When to use:

- As soon as the program begins operation.

What it shows:

- How well a program is working (e.g., how many people are participating in the program and how many people are not).

Why it is useful:

- Identifies *early* any problems that occur in reaching the target population.
- Allows programs to evaluate how well their plans, procedures, activities, and materials are working and to make adjustments before logistical or administrative weaknesses become entrenched.

Stage 3: Impact Evaluation (For details, see page 29)

When to use:

- After the program has made contact with at least one person or one group of people in the target population.

What it shows:

- The degree to which a program is meeting its intermediate goals (e.g., how awareness about the value of bicycle helmets has changed among program participants).
- Changes in the target population's knowledge, attitudes, and beliefs.

Why it is useful:

- Allows management to modify materials or move resources from a nonproductive to a productive area of the program.
- Tells programs whether they are moving toward achieving these goals.

Stage 4: Outcome Evaluation (For details, see page 32)

When to use:

- *For ongoing programs (e.g., safety classes offered each year)*: at appropriate intervals (see **When To Conduct**, page 32).
- *For one-time programs (e.g., a 6-month program to distribute car seats):* when program is complete.

What it shows:

- The degree to which the program has met its ultimate goals (e.g., how much a smoke detector program has reduced injury and death due to house fires).

Why it is useful:

- Allows programs to learn from their successes and failures and to incorporate what they have learned into their next project.
- Provides evidence of success for use in future requests for funding.

Figure 4.

Which Stage of Evaluation Are You Ready For?

To find out which stage of evaluation your program is ready for, answer the questions below. Then follow the directions provided after the answer.

Q. Does your program meet any of the following criteria?

- **It is just being planned and you want to determine how best to operate.**
- **It has some problems you do not know how to solve.**
- **It has just been modified and you want to know whether the modifications work.**
- **It has just been adapted for a new setting, population, problem, or behavior.**

Yes to any of the four criteria. Begin formative evaluation. Go to page 25.

No to **all** criteria. Read the next question.

Q. Your program is now in operation. Do you have information on who is being served, who is not being served, and how much service you are providing?

Yes. Read next question.

No. Begin process evaluation. Go to page 27. You may also be ready for impact evaluation. Read next question.

Q. Your program has completed at least one encounter with one member or one group in the target population (e.g., completed one training class). Have you measured the results of that encounter?

Yes. Read next question.

No. You are ready for impact evaluation. Go to page 29. If you believe you have had enough encounters to allow you to measure your success in meeting your overall program goals, read the next question.

Q.

- *For Ongoing Programs:*
  Has sufficient time passed and have you had contact with a sufficient number of people to allow you to measure how well the program has done in meeting its ultimate goal of reducing morbidity and mortality? See *When To Conduct*, page 32.
- *For One-Time Programs:*
  Is the program complete?

Yes. You are ready for outcome evaluation. Go to page 32.

No. Reread the questions above and, if you are still unclear, reread "Introduction" (page 21) and look at Figure 4 on page 23. If you remain uncertain, you may need to contact a professional consultant.

Figure 5.

STAGE 1: FORMATIVE EVALUATION

Description: Formative evaluation is the process of testing program plans, messages, materials, strategies, or modifications for weaknesses and strengths *before* they are put into effect. Formative evaluation is also used when an unanticipated problem occurs *after* the program is in effect.

The purpose of formative evaluation is to ensure that program materials, strategies, and activities are of the highest possible quality.

Purpose: Formative evaluation ensures that program materials, strategies, and activities are of the highest possible quality (quality assurance). During the developmental stage of a program, the purpose is to ensure that the program aspect being evaluated (e.g., a home visit to check smoke detectors) is feasible, appropriate, meaningful, and acceptable for the injury prevention program and the target population.

In the case of an unanticipated problem after the program is in effect, the purpose is to find the reason for the problem and then the solution.

When To Conduct: Conduct formative evaluation when a new program is being developed and when an existing program is 1) being modified; 2) having problems with no obvious solutions; or 3) being adapted for a new setting, population, problem, or behavior.

Target Population: Whom you ask to participate in formative evaluation depends on the evaluation's purpose. For example, if you are pilot testing materials for a new program, select people or households at random from the target population. If you want to know the level of consumer satisfaction with your program, select evaluation participants from people or households who have already been served by your program. If you want to know why fewer people than expected are taking advantage of your program, select evaluation participants from among people or households in the target population who did *not* respond to your messages.

Type of Information Produced by Formative Evaluation While a Program Is Being Developed: Whether the program being developed is surveillance or intervention, new or adapted, the formative evaluator's first concern is to answer questions similar to these:

- *Introduction:* When is the best time to introduce the program or modification to the target population?
- *Plans and Strategies:* Are the proposed plans and strategies likely to succeed?

- *Methods for Implementing Program:* Are the proposed methods for implementing program plans, strategies, and evaluation feasible, appropriate, and likely to be effective; or are they unrealistic, poorly timed, or culturally insensitive?

- *Program Activities:* Are the proposed activities suitable for the target population? That is, are they meaningful, barrier-free, culturally sensitive, and related to the desired outcome. For example, is the literacy level appropriate? Would a bicycle rodeo appeal to teenagers or would they see it as childish? Is a lottery for child safety seats acceptable or will some members of the population see it as gambling?

- *Logistics:* How much publicity and staff training are needed? Are sufficient resources (human and fiscal) available? Are scheduling and location acceptable? For example, would scheduling program hours during the normal workday make it difficult for some people in the target population to use the program?

- *Acceptance by Program Personnel:* Is the program consistent with staff's values? Are all staff members comfortable with the roles they have been assigned? For example, are they willing to distribute smoke detectors door-to-door or to participate in weekend activities in order to reach working people?

- *Barriers to Success:* Are there beliefs among the target population that work against the program? For example, do some people believe that children are safer if they are held by an adult than if they are restrained in a car seat?

Finding Solutions to Unanticipated Problems After a Program Is in Operation: If a program is already in operation but having unanticipated problems, evaluators can conduct formative evaluation to find the cause. They look at the same aspects of the program as they do during the developmental stage of the program to see 1) what is the source of the problem and 2) how to overcome the problem.

Methods: Because formative evaluators are looking for problems and obstacles, they need a format that allows evaluation participants free rein to mention whatever they believe is important. In such a case, qualitative methods (personal interviews with open-ended questions [page 38], focus groups [page 39], and participant-observation [page 39]) are best. A closed-ended quantitative method would gather information only about the topics identified in advance by program staff or the evaluator.

Occasionally, however, quantitative surveys (page 44) may be appropriate. They are useful when the purpose of evaluation is to find the level of consumer or staff satisfaction with particular aspects of the program.

How To Use Results: Well-designed formative evaluation shows which aspects of your program are likely to succeed and which need improvement. It should also show *how* problem areas can be improved. Modify the program's plans, materials, strategies, and activities to reflect the information gathered during formative evaluation.

Modify the program's plans, materials, strategies, and activities to reflect the information gathered during formative evaluation.

Ongoing Process: Formative evaluation is a dynamic process. Even after the injury prevention program has begun, formative evaluation should continue. The evaluator must create mechanisms (e.g., customer satisfaction forms to be completed by program participants) that continually provide feedback to program management from participants, staff, supervisors, and anyone else involved in the program.

Stage 2: Process Evaluation

Description: Process evaluation is the mechanism for testing whether the program's procedures for reaching the target population are working as planned.

Purpose: To count the number of people or households the program is serving, to determine whether the program is reaching the people or households it planned to serve, and to determine how many people or households in the target population the program is *not* reaching.

When To Conduct: Process evaluation should begin as soon as the program is put into action and continue throughout the life of the program. Therefore, you need to design the forms for process evaluation while the program is under development (see examples in Appendix B).

Important Factor To Consider Before Beginning Process Evaluation: The evaluator and program staff must decide whether the program should be evaluated on the basis of the number of people contacted or the number of contacts with people. The distinction is important.

For *number of people contacted*, count only *once* each person in the target population who had contact with your program regardless of how many times that person had contact.

For *number of contacts with people*, count *once* each time the program had contact with a member of the target population regardless of how many times some people had contact.

Obviously the number of *contacts with people* should be the same as, or higher than, the number of *people contacted*.

This distinction is especially meaningful when a person may receive independent value or additional benefit from each contact with the program.

Target Population: For process evaluation, the target population is the people or households that you *actually* reached, whereas your program's target population is the people or households you *want* to reach.

Methods: Keep track of all contacts with the people or households who are served by the program. If appropriate, keep track of all program-related items distributed to, or received from, the target population.

- *Direct Contacts:* One method of keeping track of direct contacts is to use simple encounter forms (see Appendix B for an example), which can be designed to collect basic information 1) about each person or household that has direct contact with the program and 2) about the nature of the contact. Using these forms, you can easily count the number of people or households served by your program, the number of items distributed during a product-distribution program, or the number of items returned to a product-loan program.

  The forms must be designed while the program is being developed and ready for use as soon as the program begins.

- *Indirect Contacts:* Not all contact with a program is direct. A program's target population may be reached directly, indirectly, or both. For example, many school-based programs provide information to schoolchildren (direct) who, in turn, take the information home to parents (indirect). Other programs train members of the target population as counselors (direct) to work with their peers in the school community (indirect). Such methods have been used by programs to promote the use of bicycle helmets. Often, a program's stated purpose is to reach community members through indirect methods. Often also, programs have an indirect effect that was not planned.

To estimate the number of people the program reaches indirectly, you could ask the people with whom the program has direct contact to keep track of their contacts (the people to whom they give the program's information or service). For this purpose, they could use a system similar to the program's system of keeping track of its direct contacts.

Sometimes, however, asking people with whom the program has direct contact to keep track of their contacts is impractical, unreliable, or both. In such a case, devise a reliable method for estimating the number of indirect contacts. For example, you could estimate that half the third graders who attended a safety training program would speak to their parents about the information given to them.

- *Items Distributed or Collected:* Example forms for use in tracking items collected from the traget population or given away during a safety product distribution campaign are in Appendix B.

How To Use Results: Use the results of process evaluation to show funding agencies the program's level of activity (i.e., the number of people or households who have received the program's service).

If process evaluation shows some unexpected problems, especially if it shows you are not reaching as many people in the target population as you expected to, do some more formative evaluation. That could include, for example, personal interviews with a random selection of people in your target population who had *not* participated in your program.

In addition, much of the information gathered during process evaluation can be used for impact and outcome evaluation when you will be calculating the effect your program has had on the target population.

Stage 3: Impact Evaluation

Description: Impact evaluation is the process of assessing the program's progress toward its goals (i.e., measuring the immediate changes brought about by the program in the target population).

Purpose: To learn about the target population's changes in knowledge, attitudes, and beliefs that may lead to changes in injury-prevention behavior. For example, evaluators might

want to know whether people are more likely to buy a bicycle helmet or smoke detector than they were before the program began. Or they might want to know whether people understand better the risks associated with not wearing seatbelts while driving.

At this stage, evaluators are *not* necessarily measuring changes in behavior (e.g., increases in the number of people *using* a bicycle helmet or smoke detector). Although information about behavior could be used to measure impact, it is a better measure of program outcome, which is the final stage of evaluation (see page 32).

To qualify for funding, programs need to incorporate evaluation—at least as far as the impact stage—into their program design.

To qualify for funding, programs need to incorporate evaluation—at least as far as the impact stage—into their program design.

When To Conduct: Take baseline measurements of the target population's knowledge, attitudes, and beliefs immediately before the first encounter with the target population (e.g., before the first training class or before any products are distributed). Begin measuring changes in knowledge, attitudes, and beliefs immediately after the first encounter (see **Methods**, beginning at the bottom of this page).

Target Population: For impact evaluation, the target population consists of people or households that received the program service.

Design: Well-designed impact evaluation has two aspects:

- It measures the baseline knowledge, attitudes, and beliefs of the target population *and* demonstrates how these change as a result of the program.

- It eliminates the possibility that any demonstrated change could be attributed to some factor outside the program. See page 51 for program designs (experimental and quasi-experimental) that control for the effect of outside influences on your program's results.

Outside Influences To Be Eliminated as Explanations for Change: The two main influences that must be eliminated as explanations for change in the participants' knowledge, attitudes, beliefs, or behaviors are *history* and *maturation*. See page 54 for a full discussion.

Methods: Measure the target population's knowledge, attitudes, beliefs, and behaviors *before* any individual or group receives the program service (baseline measurement)

and again *after* the first person or group receives the service. Compare the two measurements to find out what changes occurred as a result of your program. Be careful not to conclude that your program brought about *all* change shown by these comparisons.

Knowledge, attitudes, and beliefs are almost always measured by a survey instrument, such as a questionnaire, containing closed-ended items (e.g., multiple-choice questions). For example, you could ask each person attending a training class to complete a questionnaire before and after the class to find out how much their knowledge, attitudes, and beliefs changed as a result of the training program.

> *Tip:* Since you have a captive audience in a training class, you could also test their satisfaction with the class materials and the way the class was conducted (formative evaluation).

For information on conducting a survey, see page 44. For examples of close-ended items for survey instruments, see page 104.

Occasionally, however, knowledge, attitudes, and beliefs are assessed by direct observation. For example, an observer might check to see that seatbelts are positioned correctly or smoke detectors installed correctly. Evaluators might also observe group discussions to watch and listen for signs of participants' attitudes or beliefs (see "Participant-Observation," page 39). Observation is often more costly, less efficient, and less feasible than administering a survey instrument. For suggestions about events to observe during an evaluation, see page 89.

How To Use Results: If the results are positive, you can use the results of impact evaluation to justify continuing your program. If the results are negative, you can justify revising or discontinuing the program. Obviously, if impact evaluation shows that the program is ineffective, outcome evaluation is not necessary. Programs with positive results are likely to receive further funding; programs with negative results obviously will have a more difficult time getting funds. However, if an evaluator can show why the program was ineffective and how it could be modified to make it effective, the program may be able to justify receiving further funds.

Stage 4: Outcome Evaluation

Description: Outcome evaluation is the process of measuring whether your program met its ultimate goal of reducing morbidity and mortality due to injury.

Decisions about whether to continue funding a program often depend on the results shown by outcome evaluation. And, as you know by now, good quality outcome evaluation depends on a good design for the injury prevention program itself.

Purpose: The only difference in purpose between impact evaluation and outcome evaluation is the program effect that is measured. Impact evaluation measures changes in knowledge, attitudes, beliefs, and (possibly) preventive behaviors. Outcome evaluation measures changes in preventive behaviors and in injury-related morbidity and death.

When To Conduct: For ongoing programs (e.g., a series of fire-safety classes given each year in elementary schools), conduct outcome evaluation as soon as enough people or households have participated in the program to make outcome evaluation results meaningful. Depending on the number of children in fire-safety classes, you could conduct outcome evaluation, for example, every year, every three years or every five years.

For one-time programs (e.g., a six-month program to distribute free smoke detectors to low-income households), begin outcome evaluation as soon as the program is complete. Consider also conducting outcome evaluation, for example, a year or three years after the program is complete to find out how well the program's effects are sustained over time.

Preparation for outcome evaluation, however, begins when the program is being designed. The design of the program affects the quality of the data you will have for outcome evaluation. Furthermore, baseline data must be collected immediately before participants have their first encounters with the program.

Preparation for outcome evaluation begins when the program is being designed.

Target Population: For outcome evaluation, the target population is all the people or households that received the program service.

Methods: The methods used for measuring changes in behavior are essentially the same relatively easy methods as those used to measure changes in knowledge, attitudes, and beliefs during impact evaluation. In general, however, measuring changes in morbidity and mortality is not so easy. For example, you can measure the change in helmet-wearing behavior of children who participated in a safety training class soon after the class is over (see **Methods**, page 30, in the section on impact evaluation). Measuring the reduction in morbidity and mortality as a result of those same children's change in behavior is much more difficult.

A major cause of this difficulty is that the *number* of people who will die or suffer serious morbidity as a result of most unintentional injuries is small. In contrast, *everyone* has a certain attitude and behaves in a certain way with regard to the injury-preventive devices (e.g., bicycle helmets or smoke detectors) that your program is encouraging people to use. Therefore, documenting changes in morbidity and mortality that are directly the result of a program to reduce most unintentional injuries requires a vastly larger study population than does documenting changes in attitudes, beliefs, and behaviors.

In addition to a large study population, documenting changes in morbidity and mortality requires a long-term study, which is both expensive and time consuming.

So what to do? *Convert data on behavior change into estimates of changes in morbidity and mortality.*

When a long-term study is not feasible, you can convert the more readily accessible information on changes in behavior into estimates of changes in morbidity or mortality.

However, to do so, you must have three items of information:

- Data showing the effectiveness of the behavior in reducing morbidity or mortality (see page 119 for resources with data on various types of unintentional injury).

- Data showing the prevalence of the behavior before the program began (data on the pre-program behavior of the target population).

- Data showing the prevalence of the behavior after the program is complete (data on the post-program behavior of the target population).

Using these three sets of data, you can perform a simple series of calculations, which will estimate the number of lives saved by the program. See page 64 for full details on how to do the calculation.

> *Reiterating the Main Point:* To calculate morbidity and mortality on the basis of behavioral data, you will need information that links the behavior in question to an already-calculated risk for morbidity and mortality. Fortunately, many such data are available.

You can also convert data on behavior change into estimates of financial savings (see page 66 for details).

How To Use Results: You can use positive results of outcome evaluation as even stronger evidence than the results of impact evaluation to justify continued funding for your program. If the results (positive or negative) are likely to be of value to researchers or other programs to prevent unintentional injury, you may also be able to publish them in scientific journals.

Possible Future Study: For a behavior for which data are not available on the relationship between that behavior and risk for death or injury, you might consider doing a study to produce this information. You could justify the cost by stressing the importance of quantifying relationships between a certain behavior and risk for morbidity and mortality. The data produced by your study could then be published and used in outcome evaluation by other injury prevention programs.

SECTION 3

METHODS OF EVALUATION

Qualitative Methods

Quantitative Methods

Tables

METHODS OF EVALUATION

QUALITATIVE METHODS

INTRODUCTION

Qualitative methods allow the evaluator unlimited scope to probe the feelings, beliefs, and impressions of the people participating in the evaluation and to do so without prejudicing participants with the evaluator's own opinions.

Because qualitative methods are open-ended, they are especially valuable at the formative stage of evaluation when programs are pilot testing proposed procedures, activities, and materials. They allow the evaluator unlimited scope to probe the feelings, beliefs, and impressions of the people participating in the evaluation and to do so without prejudicing participants with the evaluator's own opinions. They also allow the evaluator to judge the intensity of people's preference for one item or another.

Qualitative methods are also useful for testing plans, procedures, and materials if a problem arises *after* they are in use. Using these methods, evaluators can usually determine the cause of any problem. Armed with knowledge about the cause, program staff can usually correct problems before major damage is done.

For example, let us say you put an advertisement in the local newspaper offering smoke detectors to low income people. Not as many people respond as you expected, and you want to know why. Conducting formative evaluation using qualitative methods will usually reveal the reason. Perhaps the advertisement cannot be understood because the language is too complex, perhaps your target population seldom reads newspapers, perhaps most people in the target population cannot go to the distribution location because it is not on a public transportation line, or perhaps the problem is due to some other factor. Whatever the cause, once you learn what the problem is, you are in a position to remedy it.

In this section, we describe three methods of conducting qualitative research: personal interviews, focus groups, and participant-observation. Each has advantages and disadvantages.

Personal Interviews

In-depth personal interviews with broad, open-ended questions are especially useful when the evaluator wants to understand either 1) the strengths and weaknesses of a new or modified program *before* it is in effect or 2) the cause of a problem should one develop *after* the program is in effect. Relatively unstructured personal interviews with members of the target population allow interviewees to express their point of view about a program's good and bad points without being prejudiced by the evaluator's own beliefs. Open-ended questions allow interviewees to focus on points of importance to them, points that may not have occurred to the evaluator. Personal interviews are particularly important when the target population differs in age, ethnicity, culture, or social background from program staff and when the program staff has a different professional background from those directing the program. Through the interview, the interviewee becomes a partner in, rather than the object of, the evaluation.[5]

Through a personal interview, the interviewee becomes a partner in, rather than the object of, the evaluation.

The interviewer's objective is to have as much of the conversation as possible generated spontaneously by the interviewee. For this reason, interviewers must avoid questions that can be answered briefly.

Personal interviews are the most appropriate form of qualitative evaluation when the subject is sensitive, when people are likely to be inhibited speaking about the topic in front of strangers, or when bringing a group of people together is difficult (e.g., in rural areas).

Personal interviews should be audiotaped and transcribed verbatim. Most commonly, evaluators analyze the results of personal interviews by looking through the transcripts for insightful comments and common themes. They then give a written report to program management. Thus, the interviewees' words become the evaluation data with direct quotes serving as useful supporting evidence of the evaluators' assessments.

Examples of open-ended questions to ask during personal interviews begin on page 76. See also the focus groups questions (page 81), many of which are suitable for personal interviews.

Focus Groups

Focus groups serve much the same function as personal interviews. The main difference is that, with focus groups, the questions are asked of groups. Ideally these groups comprise four to eight people who are likely to regard each other as equals.[6] A feeling of equality allows all members of the group to express their opinions freely. Focus groups have an advantage over individual interviews because the comments of one participant can stimulate the thoughts and ideas of another. You must conduct several focus groups because different combinations of people yield different perspectives. The more views expressed, the more likely you are to develop a good understanding of whatever situation you are investigating.

As with personal interviews, focus-group discussions should be audiotaped and transcribed verbatim. The evaluator looks for insightful comments and common threads both within groups and across groups and uses direct quotes as the evaluation data. Also as with personal interviews, evaluators analyze the data and prepare a written report for program management. Many of the same questions may be used for personal interviews and for focus groups.

On page 81 are examples of questions that might be used with focus groups during formative evaluation of a program.

Participant-Observation

Evaluation by participant-observation involves having members of the evaluation team participate (to the degree possible) in the event being observed, look at events from the perspective of a participant, and make notes about their experiences and observations. Aspects to observe include physical barriers for participants, smoothness of program operation, areas of success, and areas of weakness. Observers should be unobtrusive and ensure that their activities do not disrupt the program. They should be alert, trained in observational methods, and aware of the type of observations of greatest importance to the program evaluation.

Participant-observation is particularly valuable to the study of behavior for several reasons:

- The parties involved in a problem may not realize the effect of their actions or words on other people, or they may not be fully aware of their own reactions to particular situations.
- Unlike personal interviews or focus groups, participant-observation can produce information from people who have difficulty verbalizing their opinions and feelings.
- Problems of which participants are unaware can come to light. For example, parents may not be aware that an infant car seat is improperly installed and would therefore not report in an interview or focus group that they had difficulty understanding the instructions for installing the seat.

Unlike personal interviews or focus groups, participant-observation can produce information from people who have difficulty verbalizing their opinions and feelings.

A major disadvantage of participant-observation is that it is time consuming for the evaluator.

Examples of events to observe begin on page 89.

General Information

Who To Interview, Invite to Focus Groups, or Observe: If you are evaluating your program's methods, procedures, activities, or materials, select people similar to those your program is trying to reach. Indeed, you could even select members of the target population itself, if that is possible.

If you are conducting formative evaluation because a large group of people dropped out of the program or refused to join the program, then select people from that group to interview, observe, or invite to focus groups. They are the people most likely to provide information about aspects of the program that need correction.

Number of People To Interview, Focus Groups To Conduct, or Events To Observe: The number depends on the size and diversity of the target population.[7] The larger and more diverse the target population, the more interviews, focus groups, or observations are needed. In all cases, the more interviews, observations, or focus groups you conduct, the more likely you are to get an accurate picture of the situation you are investigating.

All qualitative evaluation must be conducted by people trained in the particular method being used.

Trained Evaluator: For several reasons, all qualitative evaluation *must* be conducted by people trained in the particular method (interview, focus group, or participant-observation) being used:

- They are experienced in asking open-ended questions (more difficult that you might think) and in probing deeper into a subject when an unexpected situation calls for such probing.
- They know how to elicit comments and keep people talking.
- They are experienced in encouraging shy people to participate in the conversation and in silencing domineering people.
- They are experienced in not showing what they feel or believe about a particular subject or about someone's response to a question.
- They do not bring their own values into the discussion.
- They recognize when the discussion has gone far afield of the evaluation's objectives.
- They know when disagreement is productive rather than counterproductive.
- Their interest in the results is more impersonal than any program staff member's interest would be.
- They know how to summarize and present the results in a meaningful way.

See Table 2 for a summary of qualitative methods of evaluation, including the advantages and disadvantages of each.

Table 2. Qualitative Methods of Evaluation

Method	Purpose	Number of People To Interview or Events To Observe	Resources Required	Advantages	Disadvantages
Personal Interviews	➢ To have individual, open-ended discussion on a range of issues. ➢ To obtain in-depth information on an individual basis about perceptions and concerns.	➢ The larger and more diverse the target population, the more people must be interviewed.	➢ Trained interviewers ➢ Written guidelines for interviewer ➢ Recording equipment ➢ A transcriber ➢ A private room	➢ Can be used to discuss sensitive subjects that interviewee may be reluctant to discuss in a group. ➢ Can probe individual experience in depth. ➢ Can be done by telephone.	➢ Time consuming to conduct interviews and analyze data. ➢ Transcription can be time-consuming and expensive. ➢ Participants are one-on-one with interviewer, which can lead to bias toward "socially acceptable" or "politically correct" responses.
Focus Groups	➢ To have an open-ended group discussion on a range of issues. ➢ To obtain in-depth information about perceptions and concerns from a group.	➢ 4 to 8 interviewees per group.	➢ Trained moderator(s) ➢ Appropriate meeting room ➢ Audio and visual recording equipment	➢ Can interview many people at once. ➢ Response from one group member can stimulate ideas of another.	➢ Individual responses influenced by group. ➢ Transcription can be expensive. ➢ Participants choose to attend and may not be representative of target population. ➢ Because of group pressure, participants may give "politically correct" responses. ➢ Harder to coordinate than individual interviews.
Participant-Observation	➢ To see firsthand how an activity operates.	➢ The number of events to observe depends on the purpose. To evaluate people's behavior during a meeting may require observation of only one event (meeting). But to see if products are installed correctly may require observation of many events (installations).	➢ Trained observers	➢ Provides firsthand knowledge of a situation. ➢ Can discover problems the parties involved are unaware of (e.g., that their own actions in particular situations cause others to react negatively). ➢ Can determine whether products are being used properly (e.g., whether an infant car seat is installed correctly). ➢ Can produce information from people who have difficulty verbalizing their points of view.	➢ Can affect activity being observed. ➢ Can be time consuming. ➢ Can be labor intensive.

Quantitative Methods

Introduction

Quantitative methods are ways of gathering objective data that can be expressed in numbers (e.g., a count of the people with whom a program had contact or the percentage of change in a particular behavior by the target population). Quantitative methods are used during process, impact, and outcome evaluation. Occasionally, they are used during formative evaluation to measure, for example, the level of participant satisfaction with the injury prevention program.

Unlike the results produced by qualitative methods, results produced by quantitative methods can be used to draw conclusions about the target population.

Unlike the results produced by qualitative methods, results produced by quantitative methods can be used to draw conclusions about the target population. For example, suppose we find that everyone in a focus group (randomly selected from bicyclists in the target population) wears a helmet while riding. We cannot then conclude that all bicyclists in the target population wear helmets. However, let's say that, instead of a focus group, we conducted a valid survey (a quantitative method) and found that 90% of respondents wear helmets while bicycling, we could then estimate that the percentage of bicyclists who wear helmets in the target population is in the 85% to 95% range.

Next we will explain four quantitative methods: counting systems, surveys, experimental designs, and quasi-experimental designs. We will also describe a method for converting quantitative data on changes in behavior by the target population into estimates of changes in morbidity and mortality (page 64) and into estimates of financial savings per dollar spent on your program (page 66).

Counting Systems

A counting system is the simplest method of quantifying your program's results and merely involves keeping written records of all events pertinent to the program (e.g, each contact with a member of the target population or each item distributed during a product-distribution program). Counting systems are especially useful for process evaluation (see page 27). Simply design and use forms on which you can record all pertinent information about each program event (see Appendix B for sample forms).

Surveys

Description: A survey is a systematic, nonexperimental method of collecting information that can be expressed numerically.

A survey is a systematic, nonexperimental method of collecting information that can be expressed numerically.

Conducting a Survey: Surveys may be conducted by interview (in person or on the telephone) or by having respondents complete, in private, survey instruments that are mailed or otherwise given to them. Which method to use is determined by the objectives of the survey. For example, if you want to survey businesses or public agencies, the telephone may be best because staff from those organizations are readily accessible by telephone. On the other hand, if you want to survey people who received a free smoke detector, personal visits to their homes may be best since many people in poor areas do not have telephones. In this example, personal visits also have the advantage of allowing you to observe whether the smoke detectors are installed and working properly.

Response rates are generally highest for personal interviews, but telephone and mail surveys allow more anonymity. Therefore, respondents are less likely to bias their responses toward what they believe to be socially acceptable or "politically correct." Telephone surveys are the quickest to conduct and are useful during the development of a program. However, households with telephones are not representative of all households. Indeed, the people we most want to reach with public health programs are often the people most likely *not* to have telephones.

Purpose of Surveys: While a program is under development, surveys have several uses:

- Surveys can identify the aspects of a program that potential users like and dislike *before* those aspects are put into effect. Such information allows you to modify the aspects that are unlikely to be successful. For example, you might ask people to rate on a scale of 1 to 5 how well they understood the instructions for installing a child's safety seat. If many people respond that they had difficulty following the instructions, then it is important to clarify the language in the instructions before distributing those instructions on a large scale.

- Surveys can gather baseline data on the knowledge, attitudes, and beliefs of the target population. For example, if your goal is to get more people to wear bicycle helmets, you can survey people in the target population, *before* the

program begins, to see how much they know about the value of bicycle helmets, what their attitudes are toward wearing bicycle helmets, and what they believe about bicycle helmets as a safety device.

- Surveys can gather baseline data on the rates at which members of the target population engage in behaviors of interest to the program. For example, if your program goal is to reduce the number of people who are injured or die in car crashes because they do not wear seatbelts, you can find out the number of people who already wear seatbelts. Having such information allows you to set a realistic goal for how much you want to increase that number.

After the program is in effect, surveys also have several uses:

- Surveys can measure the level of participants' satisfaction with the program. You can determine whether people in the target population are receiving information about the program, what the most common sources for the information are, and whether the information they are receiving is correct. With such knowledge, you can eliminate the expenses (e.g., cost of newspaper advertisements) for program aspects that are not working.

- If the program is having unexpected problems with no clear solution, surveys can often locate the source of the problem, which may then lead to the solution. For example, surveys can show you how people who do not participate in your program differ from those who do. Perhaps you will find that the people who do not participate do not have cars and therefore have difficulty getting to your location. Whatever the reason, once you know what it is, you can modify the program to remove whatever problem you discover.

- During impact evaluation, surveys can measure the effect your program is having on the target population's knowledge, attitudes, beliefs, and behaviors (i.e., how much they have changed since the program began). For example, if your bicycle-helmet program is successful, the target population's knowledge of and belief in bicycle helmets will have increased, and the attitude toward bicycle helmets will have improved.

- During impact or outcome evaluation, surveys can show how many more people report they are engaging in the behavior you are interested in (e.g., how many more people report that they fasten their seatbelts than did so before the program, or how many more people report that they have installed smoke detectors).

Selecting the Survey Population: Who to survey depends in part on the purpose of the survey. To evaluate the level of consumer satisfaction with the program, the survey population may be selected from among those who use the program. To learn about barriers that prevent people from using the program, select a survey population from among people who are eligible to use the program but do not. Before the program is in effect, select from a representative sample of the entire target population to determine what they like or dislike about the program's proposed procedures, materials, activities, and methods.

In all cases, you will need a complete list of the people or households targeted by the program. Such a list is called a *sampling frame*. From the sampling frame, you may select the people to be surveyed using statistical techniques such as *random sampling, systematic sampling, or stratified sampling*. You must use stratified sampling if you want a representative sample of both those who participate in the program and those who do not. A full discussion of sampling techniques is outside the scope of this book. However, several textbooks (e.g., *Measurement and Evaluation in Health Education and Health Promotion*[2]) can provide you with information on sampling methods.

Survey Instruments: A survey instrument is the tool used to gather the survey data. The most common one is the questionnaire. Other instruments include checklists, interview schedules, and medical examination record forms.

Methods for Administering Survey Instruments: Before designing a survey instrument, you must decide on the method you will use to administer it because the method will dictate certain factors about the instrument (length, complexity, and level of language). For example, instruments designed to be completed by the respondent without an interviewer (i.e., self-administered) must be shorter and easier to follow than those to be administered by a trained interviewer.

There are three methods for administering survey instruments: personal interview, telephone interview, or distribution (e.g., through the mail) to people who complete and return the questionnaire to the program. The advantages and disadvantages of each method are laid out in Table 3.

The best method to use depends on the purpose of the evaluation and the proposed respondents to the survey. Let's say, for example, you want to evaluate a training program. If class participants have a moderate level of education, having them complete and return a questionnaire before they leave the

classroom is clearly the least expensive and most efficient method. On the other hand, if class participants have problems reading, a questionnaire to be completed in class would not be useful, and you may need to conduct personal interviews.

Likewise, if you are evaluating a program to distribute smoke detectors in a well-defined, low-income housing area, you may need to interview. In this case, face-to-face would be better than telephone interviews, since income is an issue and some poor people do not have telephones.

Table 3. Advantages and Disadvantages of Methods of Administrating Survey Instruments

Method	Advantages	Disadvantages
Personal interviews	➤ Least selection bias: can interview people without telephones—even homeless people. ➤ Greatest response rate: people are most likely to agree to be surveyed when asked face-to-face.[8] ➤ Visual materials may be used.	➤ Most costly: requires trained interviewers and travel time and costs. ➤ Least anonymity: therefore, most likely that respondents will shade their responses toward what they believe is socially acceptable.
Telephone interviews	➤ Most rapid method. ➤ Most potential to control the quality of the interview: interviewers remain in one place, so supervisors can oversee their work. ➤ Easy to select telephone numbers at random. ➤ Less expensive than personal interviews. ➤ Better response rate than for mailed surveys.	➤ Most selection bias: omits homeless people and people without telephones. ➤ Less anonymity for respondents than for those completing instruments in private. ➤ As with personal interviews, requires a trained interviewer.
Instruments to be completed by respondent	➤ Most anonymity: therefore, least bias toward socially acceptable responses. ➤ Cost per respondent varies with response rate: the higher the response rate, the lower the cost per respondent. ➤ Less selection bias than with telephone interviews.	➤ Least control over quality of data. ➤ Dependent on respondent's reading level. ➤ Mailed instruments have lowest response rate. ➤ Surveys using mailed instruments take the most time to complete because such instruments require time in the mail and time for respondent to complete.

General Guidelines for Survey Instruments: When designing a survey instrument, keep in mind that it must appeal as much as possible to the people you hope will respond:

- Use language (in the instructions and the questions) at the reading level of the least educated people in your target population.
- Avoid abbreviations and terms that may not be easily understood by the target population.
- Keep the number of items to the minimum needed to fulfill the requirements of the survey. The more items, the less likely people are to respond.
- Make the appearance attractive. Appearance involves such factors as type font, font size, text layout, use of headings, and use of white space. The denser the text and the smaller the print, the less likely people are to respond.

Steps Involved in Designing Survey Instruments: Instrument design is a multistep process, and the steps need to be done in order.

1. *Clearly define the population you want to survey.* (See page 15, **A Description of the Target Population**.)

2. *Choose the method you will use to administer the survey.* (See page 46 for more information.)

3. *Develop the survey items meticulously.* Survey items are the questions or statements in the survey. Items that are closed-ended are easiest for respondents to complete and least subject to error. Closed-ended items are multiple-choice, scaled, or questions answerable by *yes* or *no* or by *true* or *false* (See page 104 for examples.)

4. *Put items in correct order.* Begin with the least sensitive items and gradually build to the most sensitive. Respondents will not answer sensitive questions until they are convinced of the survey's purpose and have developed a rapport with the "person behind the survey" (the person or group they believe is requesting the information).

   Demographic questions such as those about age, education, ethnicity, marital status, and income can be sensitive. For this reason, these questions should be at the end. Not only are they more likely to be answered then, but when a survey has solicited intimate or emotional information, the demographic questions draw respondents' attention away from the survey's subject matter and back to everyday activities.

Survey items should progress from general to specific, which eases respondents into a subject and therefore increases the likelihood that they will answer and do so accurately and truthfully. If the survey instrument covers several subjects (e.g., seatbelt use, speeding, and driving while intoxicated), the survey items for each subject should be grouped together, again progressing from general to specific within each group. Put the least sensitive subject first and the most sensitive last.

5. *Give the survey instrument an appropriate title.* This step is particularly important for survey instruments to be completed by the respondent, since the title is the respondent's first impression of the group collecting the information. To increase the number of responses you get, emphasize the importance of the survey in the title and show any relationship between your injury prevention program and the people you want to respond to the questionnaire. Examples of good titles are "Survey of the Health Needs of Our Community" and "Survey of Your Level of Satisfaction with Our Services."

6. *Assess the reliability of the survey instrument.* This step involves measuring the degree to which the results obtained by the survey instrument can be reproduced. Assess reliability by one of three methods: 1) determine the stability of the responses given by a respondent, 2) determine the equivalence of responses by one respondent to two different forms of the questionnaire, or 3) determine the internal consistency of the instrument, which is the degree to which all questions in the questionnaire are measuring the same thing.

Following are details on the three methods:

- *Stability* is measured by administering the survey instrument to the same person at two different times (test-retest) and comparing the responses given each time. Do not expect all traits (e.g., attitudes and beliefs) to be stable. For example, enthusiasm for wearing a bicycle helmet may wax and wane throughout a day or over weeks or seasons. Thus, measuring stability may not always be an appropriate way to assess the reliability of a survey instrument.
- *Equivalence* is measured by administering two different forms of the survey instrument (alternate forms) to the same person or set of people and comparing the responses to each. This method of measuring reliability is not often used because of the cost and difficulty of constructing one good survey instrument, let alone two equally strong forms of the instrument.

- *Internal consistency* is measured by comparing the same person's responses to various items in the survey instrument. If the answer to each item contributes to the respondent's overall score, then the answers to each question should correlate with the overall score. There are several formulas for calculating the internal consistency of a survey instrument. A discussion of those formulas is outside the scope of this book. See Anastasi's *Psychological Testing*[9] for more information.

7. *Assess the validity of the survey instrument.* Validity is the degree to which the instrument measures what it purports to measure. For example, how well data on seatbelt use gathered from questionnaires completed by respondents agree with *actual* seatbelt use reflects the questionnaire's degree of validity. Clearly, if data produced by responses to a questionnaire—in this example, the extent of self-reported seatbelt use—cannot be reproduced using a more direct method of gathering data (e.g., counting the number of people who are actually wearing seatbelts), then the questionnaire is not valid. There are three main types of validity: face validity, content validity, and construct validity.

   - *Face validity* is the degree to which the instrument *appears* to measure what it is intended to measure. Face validity is important for good rapport between interviewer (questioner) and respondent. If the interviewer informs the respondent that the survey is about safety habits, but the respondent believes it is about something else, the respondent may distrust the evaluator's intent and may refuse to answer or may not answer truthfully. Assess face validity through pilot tests (e.g., focus groups or personal interviews with a subgroup of the target population) and by having subject-matter experts review the questionnaire.

   - *Content validity* is the degree to which all relevant aspects of the topic being addressed are covered by the survey instrument. Assess content validity by having subject-matter experts review the content of the instrument.

   - *Construct validity* is the degree to which the survey instrument accurately measures the set of related traits that it is intended to measure. The easiest way to establish construct validity is to compare the results obtained using your instrument with those obtained using a related one for which validity has already been demonstrated.

If no related survey instruments exist, establish construct validity through hypothesis testing. For example, if you developed a survey instrument to determine how often people exceed the speed limit, you could hypothesize that people who most frequently exceed the speed limit are likely to have more traffic citations than people who do not often exceed the speed limit. You could then gather traffic citation data and determine whether the people identified by the survey instrument as the most frequent speeders had more citations, as hypothesized.

8. *Pilot test the survey instrument.* Before an instrument can be used on the entire target population, you must pilot test it on a group of people similar to the target population or, preferably, on a small group within the target population. The purpose is to determine whether the survey instrument is effective for use with the people who are potential respondents. The evaluator's job is to find out if any survey items are confusing, ambiguous, or phrased in language unfamiliar to the intended audience. The evaluator will also determine if certain words differ in meaning from one ethnic group to the next and if certain questions are insensitive to the feelings of many people in the target population.

   Tip: If the survey instrument is not significantly modified as a result of the pilot test (a rare event), the information gathered from the people who participated in the pilot test can be added to the information obtained from the people in the full survey.

9. *Modify.* At each step of the design, modify survey items and the survey instrument itself on the basis of information gathered at that step, particularly information gathered during the pilot test.

Many good references are available on the design of survey instruments (see "Bibliography," page 117).

Experimental and Quasi-Experimental Designs

Introduction: In this section, we discuss research designs that you can use during several stages of evaluation:

- During *formative evaluation* to pilot test particular components of a program. For example, you can determine which of several advertisements is most effective in getting

people to participate in your program or which of several media messages is best at making people aware of your program. By knowing which advertisement or message is most effective, you can conserve resources by using them only for the items you know in advance are most likely to work.

- During *impact evaluation* to measure how well a program is influencing knowledge, attitudes, and beliefs. For example, you can measure how much participants' awareness of the hazards of driving without a seatbelt has increased from what it was before a program to increase seatbelt use began.

- During *outcome evaluation* to measure how well a program met its overall goal. For example, you can measure how many more people are wearing seatbelts than before the program began and, as a consequence, how many lives have been saved and injuries prevented.

How you operate your program will be influenced by how you plan to evaluate it. If you use an experimental or quasi-experimental design, impact and outcome evaluation will be a breeze because, in effect, you will be operating and evaluating the program at the same time.

If you use an experimental or quasi-experimental design, impact and outcome evaluation will be a breeze because, in effect, you will be operating and evaluating the program at the same time.

Experimental Designs: The best designs for impact and outcome evaluation are experimental designs. Evaluation with an experimental design produces the strongest evidence that a program contributed to a change in the knowledge, attitudes, beliefs, behaviors, or injury rates of the target population.

The key factor in experimental design is *randomization*: evaluation participants are randomly assigned to one of two or more groups. One or more groups will receive an injury intervention, and the other group(s) will receive either no intervention or a placebo intervention. The effects of the program are measured by comparing the changes in the various groups' knowledge, attitudes, beliefs, behaviors, or injury rates.

Randomization ensures that the various groups are as similar as possible, thus allowing evaluators of the program's impact and outcome to eliminate factors *outside* the program as reasons for changes in program participants' knowledge, attitudes, beliefs, behavior, or injury rates. See "Factors To Be Eliminated as Contributors to Program Results" (page 54) for a full discussion.

Difficulties with Experimental Designs: Although experimental designs are ideal for program evaluation, they are often difficult—sometimes impossible—to set up. The difficulty may be due to logistical problems, budgetary limitations, or political circumstances.

To demonstrate the difficulties, let us consider the example of introducing a curriculum on bicycle safety for third graders at a certain school. Selecting children at random to participate in the program would cause many problems, including the following:

- *Logistical Problems:* The program could not be administered to children in their regular classroom, since (with randomization) not all children in a classroom would be assigned to participate.
- *Budgetary Problems:* Costs would increase if an extra teacher were required to administer the program while other teachers maintained their regular schedules.
- *Political Problems:* Parents might complain if their children were not selected for the "special program." And if, as a result of parent complaints, all children had to participate in the "special program," costs would increase and the value of randomization would be lost.

In addition, evaluation of the program's effectiveness would be compromised if children in the safety class shared information with the children who were not in the safety class.

Another difficulty with experimental designs is that participants must give their *informed consent*. People who willingly agree to participate in a program in which they *may not* receive the injury intervention are probably different from people in the general population. Therefore, program effects shown through evaluation involving randomized studies may not be generalizable (i.e., they may not reflect the probable effects for all people).

For example, let us suppose you want to test how effective a bicycle rodeo is at getting bicyclists to wear helmets. You ask a random sample of 500 children who do not own bicycle helmets to attend a bicycle rodeo you have organized for the following Saturday morning. Let's say, 300 agree to go. The 200 who do not agree are probably different from the 300 who do agree: perhaps the 200 who do not agree have other activities on Saturday morning (if they are poor, they may work; if they are rich, they may go horseback riding), or they may be rebellious and refuse to listen to adults, or they may believe bicycle helmets and bicycle rodeos are not "cool," or

they may have some other reason. Whatever the reason, it makes those who refuse to participate in the study different from those who agree. And because of that difference, the results of your study will not be generalizable to the whole population of children who do not wear bicycle helmets.

Quasi-Experimental Designs: Because of the difficulties with experimental designs, programs sometimes use quasi-experimental designs. Such designs do not require that participants be randomly assigned to one or another group. Instead, the evaluator selects a whole group (e.g., a third-grade class in one school) to receive the injury intervention and another group (e.g., the third-grade class in a different school) as the comparison or control group.

As an alternative, if a suitable comparison group cannot be found, the evaluator could take multiple measurements of the intervention group *before* providing the intervention.

When using quasi-experimental designs with comparison groups, evaluators must take extra care to ensure that the intervention group is similar to the comparison group, and they must be able to describe the ways in which the groups are *not* similar.

Factors To Be Eliminated as Contributors to Program Results

Events aside from the program can produce changes in the knowledge, attitudes, beliefs, and behaviors of your program's target population, thus making your program *seem* more successful than it actually was. Therefore, anyone evaluating an injury prevention program's success must guard against assuming that all change was produced by the program. Experimental designs *minimize* (i.e., decrease to the least possible amount) the effects of outside influences on program results; quasi-experimental designs reduce those effects.

Evaluators must guard against the effects of "history" and "maturation" on program results.

The two main factors evaluators must guard against are *history* and *maturation*.

History: What may *seem* like an effect produced by your program, *an apparent impact*, may often be more accurately attributed to *history* if the people who participate in your program are different from those who do not. For example, suppose you measured bicycle-helmet use among students at a school that had just participated in your injury-prevention program and also at a school that did not participate. Let us

say that more students wore helmets at the school with your program. You have not demonstrated that your program was the reason for difference in helmet use unless you can show that the students at the school with the program did not wear helmets any more frequently *before* the bicycle-helmet program began than did the students at the school without the program. In other words, you must show that the students at the school with the program did not have a *history* of wearing helmets more often than did the students at the school without the program.

Maturation: Sometimes events outside your program cause program participants to change their knowledge, attitudes, beliefs, or behavior *while the program is under way*. Such a change would be due to *maturation*, not to the program itself. For example, suppose you measured occupant-restraint use by the 4- and 5-year-olds who attended a year-long Saturday safety seminar, both when they began the seminar and when they completed it. Let us say that the children used their seatbelts more frequently after attending the program. You have not demonstrated that the program was effective unless you can also show that seatbelt use by a similar group of 4- and 5-year-olds did not increase just as much simply as a result of other events (e.g., the children's increased manual dexterity due to development, exposure to a children's television series about using seatbelts which ran at the same time as the seminar, or a new seatbelt law that went into effect during the course of the seminar).

Schematics for Experimental and Quasi-Experimental Designs

Introduction: The steps involved in the various experimental and quasi-experimental designs are presented verbally and then in schematic form. In each schematic, we use the same symbols:

R	=	Randomization
$\mathbf{O_1}$	=	The first, or baseline, observation (e.g., results of a survey to measure the knowledge, attitudes, beliefs, behaviors, or injury rates of the target population)
$\mathbf{O_2}$	=	The second observation ($\mathbf{O_3}$ = the third, etc.)
X	=	Intervention
P	=	Placebo (usually in parenthesis to indicate that a placebo may or may not be used)

The schematic for each intervention and comparison group is shown on a separate line. For example,

$$\mathbf{O_1 \quad X \quad O_2}$$

means that there is only one group **(one line)**, that the group is observed for a baseline measurement **(O_1)**, provided with the intervention **(X)**, and observed again **(O_2)** to measure any changes.

Another example:

$$\begin{array}{lll} \mathbf{RO_1} & \mathbf{X} & \mathbf{O_2} \\ \mathbf{RO_1} & \mathbf{(P)} & \mathbf{O_2} \end{array}$$

means that people are randomly assigned **[R]** to one of two groups **[two lines]**. Both are observed for baseline measurements **[O_1]**. One is provided with the injury intervention **[X]**; the other may or may not get a placebo intervention **[(P)]**. Both groups are observed again **[O_2]** for any change.

> *Definition of Placebo:* A placebo is a service, activity, or program material (e.g., a brochure) that is similar to the intervention service, activity, or material but without the characteristic of the intervention that is being evaluated. For example, to test the effectiveness of the content of a brochure about the value of installing smoke detectors, the intervention group will be given the brochure to read and discuss with the evaluator and the comparison group might be given a brochure on bicycle helmets to read and discuss with the evaluator.

To ensure that the placebo conditions are comparable with those of the intervention, evaluators should give the same amount of time and attention to the comparison group as they give to the intervention group.

Examples of Experimental Designs

Pretest-Posttest-Control Group Design: Scientists often call this design a *true experiment* or a *clinical trial*. These are the steps involved:

1. Recruit people for the evaluation.

2. Randomly assign each person **[R]** to one of two groups: one group will receive the injury intervention **[X]** and the other will not **[(P)]**. To select at random, use a computer-generated list of random numbers, a table of random numbers (found at the back of most books on basic statistics), or the toss of a coin.

3. Observe (measure) each group's knowledge, attitudes, beliefs, behaviors, injury rate, or any other characteristics of interest **[O_1]**. You could use a survey (page 44), for example, to make this measurement.

4. Provide the program service (the intervention) **[X]** to one group and no service or a placebo service **[(P)]** to the other group.

5. Again, observe (measure) each group's knowledge, attitudes, beliefs, behaviors, injury rates, or whatever other characteristic you measured before providing the program service **[O_2]** .

The schematic for the pretest-posttest-control group design is as follows:

RO_1	**X**	**O_2**
RO_1	**(P)**	**O_2**

The effect of the program is the difference between

the change from pretest **[O_1]** to posttest **[O_2]** for the intervention **[X]** group

and

the change from pretest **[O_1]** to posttest **[O_2]** for the comparison **[(P)]** group.

To clarify, let's take a hypothetical example of a study you might conduct during formative evaluation. Suppose you want to pilot test a proposed brochure designed to increase people's awareness that working smoke detectors save lives.

1. Select a group of people at random from the target population. This group is your study [evaluation] population.

2. Randomly assign each person in the study population either to the intervention group or to the comparison group.

3. Test each group to see what the members know about smoke detectors.

4. Decide whether to give a placebo to the comparison group.

5. Show the proposed brochure on smoke detectors only to intervention group members and allow them time to study it. If a placebo is used, show a brochure, perhaps on bicycle helmets, to the comparison group members and allow them to study it. Give the same amount of time and attention to each group.

6. To see if their awareness has increased, test each group again to measure how much they now know about smoke detectors.

Unless the proposed brochure is a dud, the intervention group's awareness of the benefits of smoke detectors will increase. However, the comparison group's test scores might also increase because of the placebo effect. For example, the comparison group might develop a rapport with the evaluators and want to please them, thus causing group members to put more thought into their responses during the second observation than they did during the first. In addition, just completing the survey at the first observation may cause them to think or learn more about smoke detectors and give better answers during the second observation.

The effect of the brochure is the difference between the change (usually increase) in the intervention group's awareness and the change (if any) in the comparison group's awareness.

Variations on the Pretest-Posttest-Control Group Design: There are several variations on the pretest-posttest-control group design.

The *pretest-posttest-control group-followup design* is used to determine whether the effect of the program is maintained over time (e.g., whether people continue to wear seatbelts months or years after a program to increase seatbelt use is over). This design involves repeating the posttest at scheduled intervals. The schematic for this design is as follows:

RO$_1$	**X**	**O$_2$**	**O$_3$**	**O$_4$**
RO$_1$	**(P)**	**O$_2$**	**O$_3$**	**O$_4$**

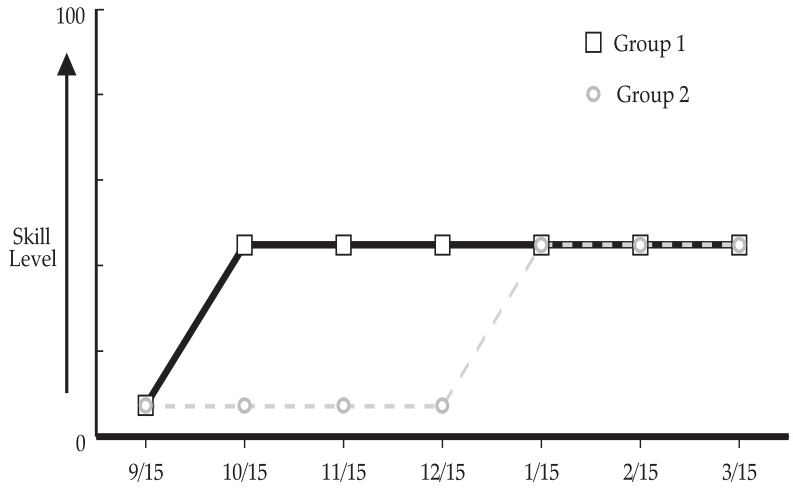
For example, suppose you want to test the effectiveness of counseling parents about infant car seats when parents bring their infants to a pediatrician for well-child care. First, select a target population for the evaluation (e.g., all the parents who seek well-child care during a given week). Then, observe (measure) the target population's use of safety seats **[O$_1$]**. Next, randomly assign some parents to receive counseling about car safety seats **[X]** and the remaining parents to receive a placebo (e.g., counseling on crib safety) **[P]**. At regular intervals after the counseling sessions, observe each group's use of infant car seats to see how well the effect of the program is maintained over time (let's say, 3 months **[O$_2$]**, 6 months **[O$_3$]**, and 9 months **[O$_4$]**) .

The ***cross-over design*** is used when everyone eligible to participate in a program must receive the intervention. Again, participants are randomly divided into two groups. Both groups are tested, but only one receives the intervention. At regular intervals, both groups are observed to see what changes (if any) have occurred in each group. After several observations, the second group receives the intervention, and both groups continue to be observed at regular intervals. Below is an example schematic for this design:

$$\begin{array}{lllllllll} \mathbf{RO_1} & \mathbf{X} & \mathbf{O_2} & \mathbf{O_3} & \mathbf{O_4} & & \mathbf{O_5} & \mathbf{O_6} & \mathbf{O_7} \\ \mathbf{RO_1} & & \mathbf{O_2} & \mathbf{O_3} & \mathbf{O_4} & \mathbf{X} & \mathbf{O_5} & \mathbf{O_6} & \mathbf{O_7} \end{array}$$

A program is effective if the effect being measured (e.g., increase in knowledge) changes for Group 1 after the first observation and for Group 2 after the fourth observation.

For example, suppose you wanted to evaluate whether children who took a fire-safety class presented by the fire department had better fire-safety skills than children who did not take the class. To conduct such an evaluation you could, for example, test the fire-safety skills of *all* the children in the third grade of the local elementary school, then randomly select *half* of the children (Group 1) to attend the fire-safety class on September 15. You would test the fire-safety skills of *all* the children again on, say, October 15, November 15, and December 15. In January the other half of the class (Group 2) would attend the fire-safety class. You would again test the fire-safety skills of *all* the children on January 15, February 15, and March 15. If the class were to increase the children's fire-safety skills, the results of evaluation might look something like this.



The *Solomon four-group design* is useful when the act of measuring people's pre-program knowledge, attitudes, beliefs, or behaviors (getting baseline measurements) may affect the program's goals in one or both of the following ways:

- *People may change their behavior as a result of being questioned about it*. For example, simply asking people how often they fasten their seatbelts may remind them to do so, thereby increasing the use of seatbelts even *before* any program to increase seatbelt use begins.

- *People's interest in a subject may increase simply because they are questioned about it*. Such an increase would affect the program's outcome. For example, simply being questioned about smoke detectors may prime program participants to be more receptive to receiving information about them during a program to prevent house fires.

To compensate for those possibilities, this design expands the pretest-posttest-control group design from two groups (one intervention and one control) to four groups (two intervention and two control). To separate the effect of getting a baseline measurement from the effect produced by the program, the evaluator takes baseline measurements of only one intervention and one control group. The four groups are distinguished from one another as shown below:

Group 1: Provides baseline measurement and receives the intervention.

[RO_1 X O_2]

Group 2: Provides baseline measurement and receives nothing or a placebo.

[RO_1 (P) O_2]

Group 3: Provides *no* baseline measurement and receives the intervention.

[R X O_2]

Group 4: Provides *no* baseline measurement and receives nothing or a placebo.

[R (P) O_2]

Since the only difference between Groups 2 and 4 is that Group 2 provided a baseline measurement and Group 4 did not, the evaluator can compare the posttest results **(O_2)** of Group 2 with those of Group 4 to determine the effect of taking a baseline observation **(O_1)**.

Similarly, since the only difference between Group 1 and Group 3 is whether they provided a baseline measurement, evaluators can compare their posttest results **(O_2)** to determine whether providing a baseline measurement primed program participants to be more interested in the program's information, thus increasing the program's effectiveness.

The schematic for the Solomon four-group design is as follows:

RO_1	**X**	**O_2**
RO_1	**(P)**	**O_2**
R	**X**	**O_2**
R	**(P)**	**O_2**

Unfortunately, however, since this variation increases the number of people required for study, it also increases the study's cost, time, and complexity. As a result, people who are willing to participate in an evaluation with this design may be even less representative of the general population than people who would participate in an evaluation with a less complex, randomized design.

Examples of Quasi-Experimental Designs

Here are some examples of quasi-experimental designs. These are useful when a randomized (experimental) design is not possible:

Nonequivalent Control Group Design: Sometimes it is difficult to introduce an injury-prevention program to some people and not to others (e.g., it is impossible to be sure that a radio campaign will reach only certain people in a town and not others). In such a case, the nonequivalent control group design is useful. It is similar to the pretest-posttest-control group design except that individual participants are not randomly assigned to separate groups. Instead an entire

group is selected to receive the program service and another group not to receive it. For example, a radio campaign could be run in one town but not in a similar town some distance away.

For this example, it is important to select two groups that are well separated geographically in order to reduce the likelihood that the effect of the injury intervention will spill over to the people who are not to receive the intervention. As the name of the design indicates, without randomization the groups will never be equivalent; however, they should be as similar as possible with respect to factors that could affect the impact of the program.

As with the pretest-posttest-control group design, pretest each group **[O_1]**; the result of the pretest shows the degree to which the two groups are not equivalent. Next, provide the intervention to one group **[X]** and a placebo or nothing **[(P)]** to the other. Then posttest each group **[O_2]**.

> *Note:* The evaluator must look at *history*, in particular, as a possible way in which the two groups are not equivalent. See page 54 for a discussion of *history* as an explanation for change.

The schematic for this design is as follows:

O_1	**X**	**O_2**
O_1	**(P)**	**O_2**

Time Series Design: Sometimes it is impossible to have a control group that is even marginally similar to the intervention group (e.g., when a state program wants to evaluate the effect of a new state law). Although other states may be willing to act as comparison groups, finding a willing state that is similar with respect to legislation, population demographics, and geography is not easy. Furthermore, it is difficult to control the collection of evaluation data by a voluntary collaborator and even more difficult to provide funding to the other state.

The time series design attempts to control for the effects of *maturation* when a comparison group cannot be found. *Maturation* is the effect that events outside the program have on program participants *while the program is under way*. See page 55 for a full discussion on maturation.

To minimize the effect of maturation on program results, take multiple measurements (e.g., **O_1** through **O_4**) of program participants' knowledge, attitudes, beliefs, or behaviors

before an injury-prevention program begins and enter those measurements into a computer. Then, using special software, you can predict the future trend of those measurements were the program not to go into effect. After the program is over, again take multiple measurements (e.g., $\mathbf{O}_5$ through $\mathbf{O}_8$) of program participants' knowledge, attitudes, beliefs, or behaviors to determine how much the actual post-program trend differs from the trend predicted by the computer.

If the actual trend in participants' knowledge, attitudes, beliefs, or behaviors during the course of the program is statistically different from the computer-predicted trend, then you can conclude that the program had an effect. The major disadvantage to this design is that it does not completely rule out the effect of outside events that occur while the program is under way. For example, this design would not separate the effect of a new law requiring bicyclists to wear helmets from the effect of increased marketing by helmet manufacturers. Although this design cannot eliminate the effects of outside events, it does limit them to those that are introduced simultaneously with the injury-prevention program.

The schematic for this design is as follows:

$$\mathbf{O}_1\ \mathbf{O}_2\ \mathbf{O}_3\ \mathbf{O}_4\ \mathbf{X}\ \mathbf{O}_5\ \mathbf{O}_6\ \mathbf{O}_7\ \mathbf{O}_8$$

Multiple Time Series Design: This design combines the advantages of the nonequivalent control group design (page 61) with those of the time series design (page 62): the effects of *history* on program results are reduced by taking multiple baseline measurements, and the effects of *maturation* are reduced by the combined use of 1) a comparison group and 2) predicted trends in baseline measurements. As with the nonequivalent control-group design, a disadvantage of this design is that the groups are not strictly equivalent and may be exposed to different events that could affect results. The schematic for this design is as follows:

$$\mathbf{O}_1\ \mathbf{O}_2\ \mathbf{O}_3\ \mathbf{O}_4\ \mathbf{X}\ \mathbf{O}_5\ \mathbf{O}_6\ \mathbf{O}_7\ \mathbf{O}_8$$

$$\mathbf{O}_1\ \mathbf{O}_2\ \mathbf{O}_3\ \mathbf{O}_4\ \quad\ \mathbf{O}_5\ \mathbf{O}_6\ \mathbf{O}_7\ \mathbf{O}_8$$

Converting Data on Behavior Change into Data on Morbidity and Mortality

You can convert data on changes in the behavior your program was designed to modify into estimates of changes in morbidity and mortality if you know the *effectiveness* of the behavior in reducing morbidity and mortality.

As an example, let us suppose your program was designed to increase seatbelt use. Let us also suppose that you counted the number of people wearing seatbelts at a random selection of locations around your city both before and after the program. You found that 20% more people in large cars and 30% more people in small cars are wearing seatbelts after the program than before.

To convert that 20% increase in seatbelt use (for people in large cars) to a decrease in deaths and injuries, you will need two sets of information:

- The difference between a traveler's likelihood of (risk for) injury or death while wearing a seatbelt and while not wearing a seatbelt.
- The number of deaths and number of injuries sustained by people involved in car crashes before the program began.

In our example, both sets of information are available.

- Boehly and Lombardo (cited in *Risk Factor Update Project: Final Report,*[10] p. IV-79) showed the relative risk for death or moderate-to-severe injury for people traveling under four conditions (see Table 4).
- Statistics on deaths and injuries due to motor vehicle crashes are available, by state and by county, from the National Highway Traffic Safety Administration's Fatality Analysis Reporting System.

Table 4. Relative Risk for Death or Moderate-to-Severe Injury in a Car Crash[10]

	Relative Risk	
Car Size	**Seatbelt Buckled**	**Seatbelt Unbuckled**
Large >3,000 lbs	1.0	2.3
Small ≤3,000 lbs	2.1	5.0

Let's say, for our example, that 125 people were severely injured or died in large cars and 500 in small cars during the year before the program began. Now the calculation:

1. Subtract the risk ratio for people *wearing* seatbelts in large cars (1.0) from the risk ratio for people *not wearing* seatbelts in large cars (2.3):

2.3 - 1.0 = 1.3

The result (1.3) is the amount of risk ratio that is attributable to *not wearing* seatbelts

2. Divide this difference (1.3) by the total risk ratio for people not wearing seatbelts (2.3):

1.3 ÷ 2.3 = 0.565

3. Express the result as a percentage:

0.565 x 100 = 56.5%

This calculation tells us that, when riding in a large car, people reduce their risk for injury or death by 56.5% if they buckle their seatbelts.

4. Multiply the percentage of decreased risk (56.5%) by the increase in the percentage of people wearing seatbelts in large cars (in our example, 20%):

56.6% x 20% = 0.566 x 0.20 = 0.1132 = 11.3%

This calculation shows that injuries and deaths are reduced by 11.3% among people in large cars when 20% more of them buckle their seatbelts.

5. Multiply the percentage of decreased risk in large cars (11.3%) by the number of injuries and deaths in large cars (in our example, 125):

11.3% x 125 = 0.113 x 125 = 14.125

This calculation shows that 14 fewer people will die or be seriously injured as a result of a 20% increase in seatbelt use by people traveling in large cars.

6. Repeat the same series of calculations for people traveling in small cars.

7. Add the numbers for large cars and for small cars to determine the total number of lives saved.

Converting Data on Behavior Change into Data on Cost Savings

To convert data on behavior change (e.g., increased seatbelt use) into estimates of financial savings per dollar spent on your program, you can do the same set of calculations as those used to convert data on behavior change into estimates of changes in morbidity and mortality (page 64). Then multiply the number of deaths and injuries prevented by the cost associated with deaths and injuries, and divide by the total cost of the program. For example, if your program to increase seatbelt use produces an estimate that it saved 14 lives during the previous year, multiply 14 by the average cost-per-person associated with a death due to injuries sustained in a car crash, then divide the result by the total cost of the program.

Summary of Quantitative Methods

Quantitative methods of evaluation allow you to express the results of your activities or program in numbers. Such results can be used to draw conclusions about the effectiveness of the program's materials, plans, activities, and target population. Table 5 lists the quantitative methods we have discussed in this chapter and the purpose of each one.

Table 5. Quantitative Methods Used in Evaluation

Method	Purpose
Counting systems	➤ To record the number of contacts with program participants and with people outside the program. ➤ To record the number of items a program distributes or receives.
Surveys	➤ To measure people's knowledge, attitudes, beliefs, or behaviors.
Experimental studies	➤ To minimize the effect of events outside the program on the assessment of a program's effectiveness.
Quasi-experimental studies	➤ To reduce the effect of events outside the program on the assessment of a program's effectiveness *when experimental studies are impractical*.
Converting data on behavior change into data on morbidity and mortality	➤ To estimate the number of deaths or injuries prevented as a result of program participants changing their behavior.
Converting data on behavior change into data on cost savings	➤ To estimate the financial savings per dollar spent on your program.

REFERENCES

1. Deniston OL, Rosenstock IM. "Evaluating Health Programs," *Public Health Rep* 1970; 85(9):835–40.

2. Green LW, Lewis FM. *Measurement and Evaluation in Health Education and Health Promotion*. Palo Alto, CA: Mayfield; 1986.

3. National Cancer Institute. *Making Health Communication Programs Work*. Bethesda, MD: National Institutes of Health; 1992. NIH Publication No. 92–1493.

4. National Center for Injury Prevention and Control (NCIPC). *Ten Leading Causes of Death*. Atlanta: NCIPC, Centers for Disease Control and Prevention; 1996.

5. Rubin HJ, Rubin IS. *Qualitative Interviewing: The Art of Hearing Data*. Thousand Oaks, CA: Sage; 1995.

6. Krueger RA. *Focus Groups: A Practical Guide for Applied Research*. 2nd ed. Thousand Oaks, CA: Sage; 1994.

7. Patton MQ. *Qualitative Evaluation and Research Methods*. 2nd ed. Beverly Hills, CA: Sage; 1990.

8. Erdos PL. *Professional Mail Surveys*. New York: McGraw-Hill; 1970.

9. Anastasi A. *Psychological Testing*. 6th ed. New York: MacMillan; 1988.

10. Breslow L, Fielding J, Afifi AA, et al. *Risk Factor Update Project: Final Report*. Atlanta: Centers for Disease Control and Prevention; 1985.