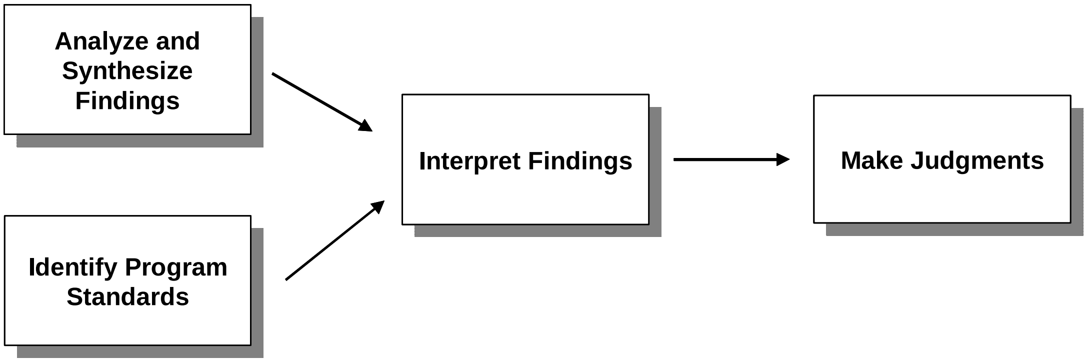
Step 5: Justify Conclusions

Whether your evaluation is conducted to show program effectiveness, help improve the program, or demonstrate accountability, you will need to analyze and interpret the evidence gathered in Step 4. Step 5 encompasses analyzing the evidence, making claims about the program based on the analysis, and justifying the claims by comparing the evidence against stakeholder values.

Why Is It Important to Justify Conclusions?

Why isn't this step called "analyze the data"? Because as central as data analysis is to evaluation, evaluators know that the evidence gathered for an evaluation does not necessarily speak for itself. As the figure below notes, conclusions become justified when analyzed and synthesized findings ("the evidence") are interpreted through the "prism" of values ("standards") that stakeholders bring, and then judged accordingly. Justification of conclusions is fundamental to utilization-focused evaluation. When agencies, communities, and other stakeholders agree that the conclusions are justified, they will be more inclined to use the evaluation results for program improvement.



The complicating factor, of course, is that different stakeholders may bring different and even contradictory standards and values to the table. As the old adage states, "where you stand depends on where you sit." Fortunately for those using the CDC Framework, the work of Step 5 benefits from the efforts of the previous steps: Differences in values and standards will have been identified at the during stakeholder engagement in Step 1. Those stakeholder perspectives will also have been reflected in the program description and evaluation focus.

Analyzing and Synthesizing The Findings

Data analysis is the process of organizing and classifying the information you have collected, tabulating it, summarizing it, comparing the results with other appropriate information, and presenting the results in an easily understandable manner. The five steps in data analysis and synthesis are straightforward:

- Enter the data into a database and check for errors. If you are using a surveillance system such as BRFSS or PRAMS, the data have already been checked, entered, and tabulated by those conducting the survey. If you are collecting data with your own instrument, you will need to select the computer program you will use to enter and analyze the data, and determine who will enter, check, tabulate, and analyze the data.

- Tabulate the data. The data need to be tabulated to provide information (such as a number or %) for each indicator. Some basic calculations include determining
  - The number of participants
  - The number of participants achieving the desired outcome
  - The percentage of participants achieving the desired outcome.
- Analyze and stratify your data by various demographic variables of interest, such as participants' race, sex, age, income level, or geographic location.
- Make comparisons. When examination of your program includes research as well as evaluation studies, use statistical tests to show differences between comparison and intervention groups, between geographic areas, or between the pre-intervention and post-intervention status of the target population.
- Present your data in a clear and understandable form. To interpret your findings and make your recommendations, you must ensure that your results are easy to understand and clearly presented. Data can be presented in tables, bar charts, pie charts, line graphs, and maps.

In evaluations that use multiple methods, patterns in evidence are detected by isolating important findings (analysis) and combining different sources of information to reach a larger understanding (synthesis).

Setting Program Standards for Performance

"Program standards" as the term is used here—and not to be confused with the four evaluation standards discussed throughout this document—are the "benchmarks" that will be used to judge program performance. They reflect stakeholders' values about the program and are fundamental to sound evaluation. The program and its stakeholders must articulate and negotiate the values that will be used to consider a program "successful," "adequate," or "unsuccessful." Possible standards that might be used in determining these benchmarks:

- Needs of participants
- Community values, expectations, and norms
- Program mission and objectives
- Program protocols and procedures
- Performance by similar programs
- Performance by a control or comparison group
- Resource efficiency
- Mandates, policies, regulations, and laws
- Judgments of participants, experts, and funders
- Institutional goals
- Social equity
- Human rights.

When stakeholders disagree about standards/values, it may reflect differences in which outcomes are deemed most important. Or, stakeholders may agree on outcomes but disagree on the *amount* of progress on an outcome necessary to judge the program a success. This threshold for each indicator, sometimes called a “benchmark” or “performance indicator,” is often based on an expected change from a known baseline. For example, all CLPP stakeholders may agree that reduction in EBLL for program participants and provider participation in screening are key outcomes to judge the program a success. But, do they agree on how *much* of an EBLL decrease must be achieved for the program to be successful, or how *many* providers need to undertake screening of children for the program to be successful? In Step 5, you will negotiate consensus on these standards and compare your results with these performance indicators to justify your conclusions about the program. Performance indicators should be achievable but challenging, and should consider the program’s stage of development, the logic model, and the stakeholders’ expectations. Identifying and addressing differences in stakeholder values/standards early in the evaluation is helpful. If definition of performance standards is done *while* data are being collected or analyzed, the process can become acrimonious and adversarial.

Interpreting the Findings and Making Judgments

Judgments are statements about a program’s merit, worth, or significance. They are formed when findings are compared against one or more selected program standards. In forming judgments about a program:

- Multiple program standards can be applied
- Stakeholders may reach different or even conflicting judgments.

Conflicting claims about a program’s quality, value, or importance often indicate that stakeholders are using different program standards or values in making their judgments. This type of disagreement can prompt stakeholders to clarify their values and reach consensus on how the program should be judged.

Tips To Remember When Interpreting Your Findings

- Interpret evaluation results with the goals of your program in mind.
- Keep your audience in mind when preparing the report. What do they need and want to know?
- Consider the limitations of the evaluation:
  - o Possible biases
  - o Validity of results
  - o Reliability of results
- Are there alternative explanations for your results?
- How do your results compare with those of similar programs?
- Have the different data collection methods used to measure your progress shown similar results?
- Are your results consistent with theories supported by previous research?
- Are your results similar to what you expected? If not, why do you think they may be different?

Source: US Department of Health and Human Services. Introduction to program evaluation for comprehensive tobacco control programs. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, Office on Smoking and Health, November 2001.

Illustrations from Cases

Let's use the affordable housing program to illustrate the main points of this chapter about the sources of stakeholder disagreements and how they may influence an evaluation. For example, the various stakeholders may disagree about the key outcomes for success. Maybe the organization's staff, and even the family, deem the completion and sale of the house as most important. By contrast, the civic and community associations that sponsor houses and supply volunteers or the foundations that fund the organization's infrastructure may demand that home ownership produce improvement in life outcomes, such as better jobs or academic performance. Even when stakeholders agree on the outcomes, they may disagree about the amount of progress that needs to be made on these outcomes. For example, while churches may want to see improved life outcomes just for the individual families they sponsor, some foundations may be attracted to the program by the chance to change communities as a whole by changing the mix of renters and homeowners. As emphasized earlier in the chapter, it is important to identify these values and disagreements about values early in the evaluation so that consensus can be negotiated and so that program description and evaluation design and focus reflect the needs of the stakeholders who need and will use the data.

Standards for Step 5
Justify Conclusions

Standard	Questions
Utility	Have you carefully described the perspectives, procedures, and rationale used to interpret the findings? Have stakeholders considered different approaches for interpreting the findings?
Feasibility	Is the approach to analysis and interpretation appropriate to the level of expertise and resources?
Propriety	Have the standards and values of those less powerful or those most affected by the program been taken into account in determining standards for success?
Accuracy	Can you explicitly justify your conclusions? Are the conclusions fully understandable to stakeholders?

Checklist for Justifying Your Conclusions

- □ Analyze data using appropriate techniques.
- □ Check data for errors.
- □ Consider issues of context when interpreting data.
- □ Assess results against available literature and results of similar programs.
- □ If multiple methods have been employed, compare different methods for consistency in findings.
- □ Consider alternative explanations.
- □ Use existing standards (e.g., *Healthy People 2010* objectives) as a starting point for comparisons.
- □ Compare program outcomes with those of previous years.
- □ Compare actual with intended outcomes.
- □ Document potential biases.
- □ Examine the limitations of the evaluation.

Worksheet 5
Justify Conclusions

Question		Response
1	Who will analyze the data (and who will coordinate this effort)?	
2	How will data be analyzed and displayed?	
3	Against what "standards" will you compare your interpretations in forming your judgments?	
4	Who will be involved in making interpretations and judgments and what process will be employed?	
5	How will you deal with conflicting interpretations and judgments?	
6	Are your results similar to what you expected? If not, why do you think they are different?	
7	Are there alternative explanations for your results?	
8	How do your results compare with those of similar programs?	
9	What are the limitations of your data analysis and interpretation process (e.g., potential biases, generalizability of results, reliability, validity)?	
10	If you used multiple indicators to answer the same evaluation question, did you get similar results?	
11	Will others interpret the findings in an appropriate manner?	

Step 6: Ensure Use of Evaluation Findings and Share Lessons Learned

The ultimate purpose of program evaluation is to use the information to improve programs. The purpose(s) you identified early in the evaluation process should guide the use of the evaluation results. The evaluation results can be used to demonstrate the effectiveness of your program, identify ways to improve your program, modify program planning, demonstrate accountability, and justify funding.

Additional uses include the following:

- To demonstrate to legislators or other stakeholders that resources are being well spent and that the program is effective.
- To aid in forming budgets and justify the allocation of resources.
- To compare outcomes with those of previous years.
- To compare actual outcomes with intended outcomes.
- To suggest realistic intended outcomes.
- To support annual and long-range planning.
- To focus attention on issues important to your program.
- To promote your program.
- To identify partners for collaborations.
- To enhance the image of your program.
- To retain or increase funding.
- To provide direction for program staff.
- To identify training and technical assistance needs.

What's involved in ensuring use and sharing lessons learned? Five elements are important in making sure that the findings from an evaluation are used:

- Recommendations
- Preparation
- Feedback
- Follow-up
- Dissemination

Making Recommendations

Recommendations are actions to consider as a result of an evaluation. Recommendations can strengthen an evaluation when they anticipate and react to what users want to know, and may undermine an evaluation's credibility if they are not supported by enough evidence, or are not in keeping with stakeholders' values.

Your recommendations will depend on the audience and the purpose of the evaluation (see text box). Remember, you identified many or all of these key audiences in Step 1, and have engaged many of them throughout as stakeholders. Hence, you have maximized the chances that the recommendations that you eventually make are relevant and useful to them. You know the information your stakeholders want and what is important to them. Their feedback early on in the evaluation makes their eventual support of your recommendations more likely.

Some Potential Audiences for Recommendations

- Local programs
- The state health department
- City councils
- State legislators
- Schools
- Workplace owners
- Parents
- Police departments or enforcement agencies
- Health care providers
- Contractors
- Health insurance agencies
- Advocacy groups

Illustrations from Cases

Here are some examples, using the case illustrations, of recommendations tailored to different purposes and for different audiences:

Audience: Local provider immunization program.
Purpose of Evaluation: Improve program efforts.
Recommendation: Thirty-five percent of providers in Region 2 recalled the content of the monthly provider newsletter. To meet the current objective of a 50% recall rate among this population group, we recommend varying the media messages by specialty, and increasing the number of messages targeted through journals for the targeted specialties.

Audience: Legislators.
Purpose of Evaluation: Demonstrate effectiveness.
Recommendation: Last year, a targeted education and media campaign about the need for private provider participation in adult immunization was conducted across the state. Eighty percent of providers were reached by the campaign and reported a change in attitudes towards adult immunization—a twofold increase from the year before. We recommend the campaign be continued and expanded to include an emphasis on minimizing missed opportunities of providers to conduct adult immunizations.

Audience: County health commissioners.
Purpose of Evaluation: Demonstrate effectiveness of CLPP efforts.
Recommendation: In this past year, county staff identified all homes with EBLL children in targeted sections of the county. Data indicate that only 30% of these homes have been treated to eliminate the source of the lead poisoning. We recommend that you incorporate compliance checks for the lead ordinance into the county's housing inspection process and apply penalties for noncompliance by private landlords.

Audience: Foundation funding source for affordable housing program.
Purpose of Evaluation: Demonstrate fiscal accountability.
Recommendation: For the past 5 years, the program has worked through local coalitions, educational campaigns, and media efforts to increase engagement of volunteers and sponsors, and to match them with 300 needy families to build and sell a house. More than 90% of the families are

still in their homes and making timely mortgage payments. But, while families report satisfaction with their new housing arrangement, we do not yet see evidence of changes in employment and school outcomes. We recommend continued support for the program but expansion to include an emphasis on tutoring and life coaching by the volunteers.

Preparation

Preparation refers to the steps taken to get ready to eventually use the evaluation findings. Through preparation, stakeholders can:

- Strengthen their ability to translate new knowledge into appropriate action.
- Discuss how potential findings might affect decision-making.
- Explore positive and negative implications of potential results and identify different options for program improvement.

Feedback

Feedback is the communication that occurs among everyone involved in the evaluation. Feedback, necessary at all stages of the evaluation process, creates an atmosphere of trust among stakeholders. Early in an evaluation, the process of giving and receiving feedback keeps an evaluation on track by keeping everyone informed about how the program is being implemented and how the evaluation is proceeding. As the evaluation progresses and preliminary results become available, feedback helps ensure that primary intended users and other stakeholders have opportunities to comment on evaluation decisions. Valuable feedback can be obtained by holding discussions and routinely sharing interim findings, provisional interpretations, and draft reports.

Follow-up

Although follow-up refers to the support that many users need throughout the evaluation process, in this step, in particular, it refers to the support that is needed after users receive evaluation results and begin to reach and justify their conclusions. Active follow-up can achieve the following:

- Remind users of the intended uses of what has been learned.
- Help to prevent misuse of results by ensuring that evidence is applied to the questions that were the evaluation’s central focus.
- Prevent lessons learned from becoming lost or ignored in the process of making complex or political decisions.

Dissemination: Sharing the Results and the Lessons Learned From Evaluation

Dissemination is the process of communicating evaluation procedures or lessons learned to relevant audiences in a timely, unbiased, and consistent manner. Regardless of how communications are structured, the goal for dissemination is to achieve full disclosure and impartial reporting. Planning effective communications requires

- Advance discussion of the reporting strategy with intended users and other stakeholders
- Matching the timing, style, tone, message source, vehicle, and format of information products to the audience.

Some methods of getting the information to your audience include

- Mailings
- Web sites
- Community forums
- Media (television, radio, newspaper)
- Personal contacts
- Listservs
- Organizational newsletters.

If a formal evaluation report is the chosen format, the evaluation report must clearly, succinctly, and impartially communicate all parts of the evaluation (see text box). The report should be written so that it is easy to understand. It need not be lengthy or technical. You should also consider oral presentations tailored to various audiences. An outline for a traditional evaluation report might look like this:

- **Executive Summary**
- **Background and Purpose**
    - Program background
    - Evaluation rationale
    - Stakeholder identification and engagement
    - Program description
    - Key evaluation questions/focus
- **Evaluation Methods**
    - Design
    - Sampling procedures
    - Measures or indicators
    - Data collection procedures
    - Data processing procedures
    - Analysis
    - Limitations
- **Results**
- **Discussion and Recommendations**

Tips for
Writing Your Evaluation Report

- Tailor the report to your audience; you may need a different version of your report for each segment of your audience.
- Present clear and succinct results.
- Summarize the stakeholder roles and involvement.
- Explain the focus of the evaluation and its limitations.
- Summarize the evaluation plan and procedures.
- List the strengths and weaknesses of the evaluation.
- List the advantages and disadvantages of the recommendations.
- Verify that the report is unbiased and accurate.
- Remove technical jargon.
- Use examples, illustrations, graphics, and stories.
- Prepare and distribute reports on time.
- Distribute reports to as many stakeholders as possible

Applying Standards

The three standards that most directly apply to Step 6—Ensure Use and Share Lessons Learned—are utility, propriety, and accuracy. As you use your own evaluation results, the questions presented in Table 6.1 can help you to clarify and achieve these standards.

Table 6.1
Standards for Step 6:
Ensure Use and Share Lessons Learned

Standard	Questions
Utility	• Do reports clearly describe the program, including its context, and the evaluation's purposes, procedures, and findings? • Have you shared significant mid-course findings and reports with users so that the findings can be used in a timely fashion? • Have you planned, conducted, and reported the evaluation in ways that encourage follow-through by stakeholders?
Feasibility	• Is the format appropriate to your resources and to the time and resources of the audience?
Propriety	• Have you ensured that the evaluation findings (including the limitations) are made accessible to everyone affected by the evaluation and others who have the right to receive the results?
Accuracy	• Have you tried to avoid the distortions that can be caused by personal feelings and other biases? • Do evaluation reports impartially and fairly reflect evaluation findings?

Evaluation is a practical tool that states can use to inform programs' efforts and assess their impact. Program evaluation should be well integrated into the day-to-day planning, implementation, and management of public health programs. Program evaluation complements CDC's operating principles for public health, which include using science as a basis for decision-making and action, expanding the quest for social equity, performing effectively as a service agency, and making efforts outcome-oriented. These principles highlight the need for programs to develop clear plans, inclusive partnerships, and feedback systems that support ongoing improvement. CDC is committed to providing additional tools and technical assistance to states and partners to build and enhance their capacity for evaluation.

Checklist for Ensuring That Evaluation Findings Are Used and Sharing Lessons Learned

□ Identify strategies to increase the likelihood that evaluation findings will be used.

□ Identify strategies to reduce the likelihood that information will be misinterpreted.

□ Provide continuous feedback to the program.

□ Prepare stakeholders for the eventual use of evaluation findings.

□ Identify training and technical assistance needs.

□ Use evaluation findings to support annual and long-range planning.

□ Use evaluation findings to promote your program.

□ Use evaluation findings to enhance the public image of your program.

□ Schedule follow-up meetings to facilitate the transfer of evaluation conclusions.

□ Disseminate procedures used and lessons learned to stakeholders.

□ Consider interim reports to key audiences.

□ Tailor evaluation reports to audience(s.)

□ Revisit the purpose(s) of the evaluation when preparing recommendations.

□ Present clear and succinct findings in a timely manner.

□ Avoid jargon when preparing or presenting information to stakeholders.

□ Disseminate evaluation findings in several ways.

Worksheet 6A
Communicating Results

I need to communicate to this audience		This format would be most appropriate	This channel(s) would be most effective
1			
2			
3			
4			
5			
6			

Worksheet 6B
Ensuring Follow-up

The following will follow up with users of the evaluation findings		In this manner	This support is available for follow-up
1			
2			
3			
4			
5			
6			

Glossary

Accountability: The responsibility of program managers and staff to provide evidence to stakeholders and funding agencies that a program is effective and in conformance with its coverage, service, legal, and fiscal requirements.

Accuracy: The extent to which an evaluation is truthful or valid in what it says about a program, project, or material.

Activities: The actual events or actions that take place as a part of the program.

Attribution: The estimation of the extent to which any results observed are caused by a program, meaning that the program has produced incremental effects.

Breadth: The scope of the measurement's coverage.

Case study: A data collection method that involves in-depth studies of specific cases or projects within a program. The method itself is made up of one or more data collection methods (such as interviews and file review).

Causal inference: The logical process used to draw conclusions from evidence concerning what has been produced or "caused" by a program. To say that a program produced or caused a certain result means that, if the program had not been there (or if it had been there in a different form or degree), then the observed result (or level of result) would not have occurred.

Comparison group: A group not exposed to a program or treatment. Also referred to as a control group.

Comprehensiveness: Full breadth and depth of coverage on the evaluation issues of interest.

Conclusion validity: The ability to generalize the conclusions about an existing program to other places, times, or situations. Both internal and external validity issues must be addressed if such conclusions are to be reached.

Confidence level: A statement that the true value of a parameter for a population lies within a specified range of values with a certain level of probability.

Control group: In quasi-experimental designs, a group of subjects who receive all influences except the program in exactly the same fashion as the treatment group (the latter called, in some circumstances, the experimental or program group). Also referred to as a non-program group.

Cost-benefit analysis: An analysis that combines the benefits of a program with the costs of the program. The benefits and costs are transformed into monetary terms.

Cost-effectiveness analysis: An analysis that combines program costs and effects (impacts). However, the impacts do not have to be transformed into monetary benefits or costs.

Cross-sectional data: Data collected at one point in time from various entities.

Data collection method: The way facts about a program and its outcomes are amassed. Data collection methods often used in program evaluations include literature search, file review, natural observations, surveys, expert opinion, and case studies.

Depth: A measurement's degree of accuracy and detail.

Descriptive statistical analysis: Numbers and tabulations used to summarize and present quantitative information concisely.

Diffusion or imitation of treatment: Respondents in one group get the effect intended for the treatment (program) group. This is a threat to internal validity.

Direct analytic methods: Methods used to process data to provide evidence on the direct impacts or outcomes of a program.

Evaluation design: The logical model or conceptual framework used to arrive at conclusions about outcomes.

Evaluation plan: A written document describing the overall approach or design that will be used to guide an evaluation. It includes what will be done, how it will be done, who will do it, when it will be done, why the evaluation is being conducted, and how the findings will likely be used.

Evaluation strategy: The method used to gather evidence about one or more outcomes of a program. An evaluation strategy is made up of an evaluation design, a data collection method, and an analysis technique.

Ex ante cost-benefit or cost-effectiveness analysis: A cost-benefit or cost-effectiveness analysis that does not estimate the actual benefits and costs of a program but that uses hypothesized before-the-fact costs and benefits. This type of analysis is used for planning purposes rather than for evaluation.

Ex post cost-benefit or cost-effectiveness analysis: A cost-benefit or cost-effectiveness analysis that takes place after a program has been in operation for some time and that is used to assess actual costs and actual benefits.

Executive summary: A nontechnical summary statement designed to provide a quick overview of the full-length report on which it is based.

Experimental (or randomized) designs: Designs that try to ensure the initial equivalence of one or more control groups to a treatment group by administratively creating the groups through random assignment, thereby ensuring their mathematical equivalence. Examples of experimental or randomized designs are randomized block designs, Latin square designs, fractional designs, and the Solomon four-group.

Expert opinion: A data collection method that involves using the perceptions and knowledge of experts in functional areas as indicators of program outcome.

External validity: The ability to generalize conclusions about a program to future or different conditions. Threats to external validity include selection and program interaction, setting and program interaction, and history and program interaction.

File review: A data collection method involving a review of program files. There are usually two types of program files: general program files and files on individual projects, clients, or participants.

Focus group: A group of people selected for their relevance to an evaluation that is engaged by a trained facilitator in a series of discussions designed for sharing insights, ideas, and observations on a topic of concern.

History: Events outside the program that affect the responses of those involved in the program.

History and program interaction: The conditions under which the program took place are not representative of future conditions. This is a threat to external validity.

Ideal evaluation design: The conceptual comparison of two or more situations that are identical except that in one case the program is operational. Only one group (the treatment group) receives the program; the other groups (the control groups) are subject to all pertinent influences except for the operation of the program, in exactly the same fashion as the treatment group. Outcomes are measured in exactly the same way for both groups and any differences can be attributed to the program.

Implicit design: A design with no formal control group and where measurement is made after exposure to the program.

Indicator: A specific, observable, and measurable characteristic or change that shows the progress a program is making toward achieving a specified outcome.

Inferential statistical analysis: Statistical analysis using models to confirm relationships among variables of interest or to generalize findings to an overall population.

Informal conversational interview: An interviewing technique that relies on the natural flow of a conversation to generate spontaneous questions, often as part of an ongoing observation of the activities of a program.

Inputs: Resources that go into a program in order to mount the activities successfully.

Instrumentation: The effect of changing measuring instruments from one measurement to another, as when different interviewers are used. This is a threat to internal validity.

Interaction effect: The joint net effect of two (or more) variables affecting the outcome of a quasi-experiment.

Internal validity: The ability to assert that a program has caused measured results (to a certain degree), in the face of plausible potential alternative explanations. The most common threats to internal validity are history, maturation, mortality, selection bias, regression artifacts, diffusion, and imitation of treatment and testing.

Interview guide: A list of issues or questions to be raised in the course of an interview.

Interviewer bias: The influence of the interviewer on the interviewee. This may result from several factors, including the physical and psychological characteristics of the interviewer, which may affect the interviewees and cause differential responses among them.

List sampling: Usually in reference to telephone interviewing, a technique used to select a sample. The interviewer starts with a sampling frame containing telephone numbers, selects a unit from the frame, and conducts an interview over the telephone either with a specific person at the number or with anyone at the number.

Literature search: A data collection method that involves an identification and examination of research reports, published papers, and books.

Logic model: A systematic and visual way to present the perceived relationships among the resources you have to operate the program, the activities you plan to do, and the changes or results you hope to achieve.

Longitudinal data: Data collected over a period of time, sometimes involving a stream of data for particular persons or entities over time.

Macro-economic model: A model of the interactions between the goods, labor, and assets markets of an economy. The model is concerned with the level of outputs and prices based on the interactions between aggregate demand and supply.

Main effects: The separate independent effects of each experimental variable.

Matching: Dividing the population into “blocks” in terms of one or more variables (other than the program) that are expected to have an influence on the impact of the program.

Maturation: Changes in the outcomes that are a consequence of time rather than of the program, such as participant aging. This is a threat to internal validity.

Measurement validity: A measurement is valid to the extent that it represents what it is intended and presumed to represent. Valid measures have no systematic bias.

Measuring devices or instruments: Devices that are used to collect data (such as questionnaires, interview guidelines, and observation record forms).

Micro-economic model: A model of the economic behavior of individual buyers and sellers, in a specific market and set of circumstances.

Monetary policy: Government action that influences the money supply and interest rates. May also take the form of a program.

Mortality: Treatment (or control) group participants dropping out of the program. It can undermine the comparability of the treatment and control groups and is a threat to internal validity.

Multiple lines of evidence: The use of several independent evaluation strategies to address the same evaluation issue, relying on different data sources, on different analytical methods, or on both.

Natural observation: A data collection method that involves on-site visits to locations where a program is operating. It directly assesses the setting of a program, its activities, and individuals who participate in the activities.

Non-probability sampling: When the units of a sample are chosen so that each unit in the population does not have a calculable non-zero probability of being selected in the sample.

Non-response: A situation in which information from sampling units is unavailable.

Non-response bias: Potential skewing because of non-response. The answers from sampling units that do produce information may differ on items of interest from the answers from the sampling units that do not reply.

Non-sampling error: The errors, other than those attributable to sampling, that arise during the course of almost all survey activities (even a complete census), such as respondents' different interpretation of questions, mistakes in processing results, or errors in the sampling frame.

Objective data: Observations that do not involve personal feelings and are based on observable facts. Objective data can be measured quantitatively or qualitatively.

Objectivity: Evidence and conclusions that can be verified by someone other than the original authors.

Order bias: A skewing of results caused by the order in which questions are placed in a survey.

Outcome effectiveness issues: A class of evaluation issues concerned with the achievement of a program's objectives and the other impacts and effects of the program, intended or unintended.

Outcome evaluation: The systematic collection of information to assess the impact of a program, present conclusions about the merit or worth of a program, and make recommendations about future program direction or improvement.

Outcomes: The results of program operations or activities; the effects triggered by the program. (For example, increased knowledge, changed attitudes or beliefs, reduced tobacco use, reduced TB morbidity and mortality.)

Outputs: The direct products of program activities; immediate measures of what the program did.

Plausible hypotheses: Likely alternative explanations or ways of accounting for program results, meaning those involving influences other than the program.

Population: The set of units to which the results of a survey apply.

Primary data: Data collected by an evaluation team specifically for the evaluation study.

Probability sampling: The selection of units from a population based on the principle of randomization. Every unit of the population has a calculable (non-zero) probability of being selected.

Process evaluation: The systematic collection of information to document and assess how a program was implemented and operates.

Program evaluation: The systematic collection of information about the activities, characteristics, and outcomes of programs to make judgments about the program, improve program effectiveness, and/or inform decisions about future program development.

Program goal: A statement of the overall mission or purpose(s) of the program.

Propriety: The extent to which the evaluation has been conducted in a manner that evidences uncompromising adherence to the highest principles and ideals (including professional ethics, civil law, moral code, and contractual agreements).

Qualitative data: Observations that are categorical rather than numerical, and often involve knowledge, attitudes, perceptions, and intentions.

Quantitative data: Observations that are numerical.

Quasi-experimental design: Study structures that use comparison groups to draw causal inferences but do not use randomization to create the treatment and control groups. The treatment group is usually given. The control group is selected to match the treatment group as closely as possible so that inferences on the incremental impacts of the program can be made.

Random digit dialing: In telephone interviewing, a technique used to select a sample. A computer, using a probability-based dialing system, selects and dials a number for the interviewer.

Randomization: Use of a probability scheme for choosing a sample. This can be done using random number tables, computers, dice, cards, and so forth.

Regression artifacts: Pseudo-changes in program results occurring when persons or treatment units have been selected for the program on the basis of their extreme scores. Regression artifacts are a threat to internal validity.

Reliability: The extent to which a measurement, when repeatedly applied to a given situation consistently produces the same results if the situation does not change between the applications. Reliability can refer to the stability of the measurement over time or to the consistency of the measurement from place to place.

Replicate sampling: A probability sampling technique that involves the selection of a number of independent samples from a population rather than one single sample. Each of the smaller samples is termed a replicate and is independently selected on the basis of the same sample design.

Resources: Assets available and anticipated for operations. They include people, equipment, facilities, and other things used to plan, implement, and evaluate programs.

Sample size: The number of units to be sampled.

Sample size formula: An equation that varies with the type of estimate to be made, the desired precision of the sample and the sampling method, and which is used to determine the required minimum sample size.

Sampling error: The error attributed to sampling and measuring a portion of the population rather than carrying out a census under the same general conditions.

Sampling frame: Complete list of all people or households in the target population.

Sampling method: The method by which the sampling units are selected (such as systematic or stratified sampling).

Sampling unit: The unit used for sampling. The population should be divisible into a finite number of distinct, non-overlapping units, so that each member of the population belongs to only one sampling unit.

Secondary data: Data collected and recorded by another (usually earlier) person or organization, usually for different purposes than the current evaluation.

Selection and program interaction: The uncharacteristic responsiveness of program participants because they are aware of being in the program or being part of a survey. This interaction is a threat to internal and external validity.

Selection bias: When the treatment and control groups involved in the program are initially statistically unequal in terms of one or more of the factors of interest. This is a threat to internal validity.

Setting and program interaction: When the setting of the experimental or pilot project is not typical of the setting envisioned for the full-scale program. This interaction is a threat to external validity.

Stakeholders: People or organizations that are invested in the program or that are interested in the results of the evaluation or what will be done with results of the evaluation.

Standard: A principle commonly agreed to by experts in the conduct and use of an evaluation for the measure of the value or quality of an evaluation (e.g., accuracy, feasibility, propriety, utility).

Standard deviation: The standard deviation of a set of numerical measurements (on an "interval scale"). It indicates how closely individual measurements cluster around the mean.

Standardized format interview: An interviewing technique that uses open-ended and closed-ended interview questions written out before the interview in exactly the way they are asked later.

Statistical analysis: The manipulation of numerical or categorical data to predict phenomena, to draw conclusions about relationships among variables or to generalize results.

Statistical model: A model that is normally based on previous research and permits transformation of a specific impact measure into another specific impact measure, one specific impact measure into a range of other impact measures, or a range of impact measures into a range of other impact measures.

Statistically significant effects: Effects that are observed and are unlikely to result solely from chance variation. These can be assessed through the use of statistical tests.

Stratified sampling: A probability sampling technique that divides a population into relatively homogeneous layers called strata, and selects appropriate samples independently in each of those layers.

Subjective data: Observations that involve personal feelings, attitudes, and perceptions. Subjective data can be measured quantitatively or qualitatively.

Surveys: A data collection method that involves a planned effort to collect needed data from a sample (or a complete census) of the relevant population. The relevant population consists of people or entities affected by the program (or of similar people or entities).

Testing bias: Changes observed in a quasi-experiment that may be the result of excessive familiarity with the measuring instrument. This is a potential threat to internal validity.

Treatment group: In research design, the group of subjects that receives the program. Also referred to as the experimental or program group.

Utility: The extent to which an evaluation produces and disseminates reports that inform relevant audiences and have beneficial impact on their work.

Program Evaluation Resources

Some Web-based Resources

Centers for Disease Control and Prevention: http://www.cdc.gov/eval/
Community Tool Box, University of Kansas: http://ctb.ku.edu/
Harvard Family Research Project: http://www.gse.harvard.edu/hfrp/
Innovation Network: http://innonet.org
University of Wisconsin Cooperative Extension:
- Evaluation Resources: http://www.uwex.edu/ces/pdande/
- Logic Model Course: http://www1.uwex.edu/ces/lmcourse

W.K. Kellogg Foundation: http://www.wkkf.org/Programming/Overview.aspx?CID=281

Selected Publications

Connell JP, Kubisch AC, Schorr LB, Weiss, CH. New approaches to evaluating community initiatives. New York, NY: Aspen Institute, 1995.

Fawcett SB, Paine-Andrews A, Francisco VT, Schulz J, Ritchter KP, et al. Evaluating community initiatives for health and development. In: Rootman I, Goodstadt M, Hyndman B, et al., eds. Evaluating Health Promotion Approaches. Copenhagen, Denmark: World Health Organization (Euro), 1999 (In press).

Fawcett SB, Sterling TD, Paine Andrews A, Harris KJ, Francisco VT, et al. Evaluating community efforts to prevent cardiovascular diseases. Atlanta, GA: Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, 1995.

Fetterman DM, Kaftarian SJ, Wandersman A. Empowerment evaluation: Knowledge and tools for self-assessment and accountability. Thousand Oaks, CA: Sage Publications, 1996,

Patton MQ. Utilization-focused evaluation. Thousand Oaks, CA: Sage Publications, 1997.

Rossi PH, Freeman HE, Lipsey MW. Evaluation: A systematic approach. Newbury Park, CA: Sage Publications, 1999.

Shadish WR, Cook TD, Leviton LC. Foundations of program evaluation. Newbury Park, CA: Sage Publications, 1991.

Taylor-Powell E, Steele S, Douglas M. Planning a program evaluation. Madison, WI: University of Wisconsin Cooperative Extension, 1996 (see Web-based entry on page 66).

University of Toronto, Health Communication Unit at the Center for Health Promotion. Evaluating health promotion programs (see Web-based entry on page 66).