

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).