2 Evaluation Questions and Standards of Effectiveness

A Reader's Guide to Chapter 2

Evaluation Questions
Goals and objectives; participants and effectiveness; program activities, organization, and effectiveness; economics and costs; program environment

Setting Standards: What They Are and How to Set Them

How to Set Standards
Setting standards using comparisons with other programs, experts, community data sets, and the literature; evaluation standards and economic evaluations

Evaluation Questions and Standards: Establishing a Healthy Relationship

When to Set Standards

The QSV Report: Questions, Standards, Variables

Evaluation Questions

Goals and Objectives

Evaluators use evaluation questions to guide them in gathering and analyzing data on the characteristics and merits of programs. In most evaluations, one of the evaluator's main concerns is to find out whether the program's goals and objectives have been met. The goals are usually meant to be relatively general and long-term, as shown in Example 2.1.

Example 2.1 Typical Program Goals

- For the public or the community at large
  Optimize health status
  Improve quality of life
  Foster improved physical, social, and psychological functioning
  Support new knowledge about health care
  Enhance satisfaction with health care
- For health care practitioners
  Promote research
  Enhance knowledge
  Support access to new technology and practices
  Improve the quality of care delivered
  Improve education
  Foster the delivery of efficient care
- For institutions
  Improve institutional organization, structure, and efficiency
  Optimize institutional ability to deliver accessible high-quality care and superior education
- For the health care system
  Expand capacity to provide high-quality care
  Support the efficient provision of care
  Ensure respect for the health care needs of all citizens

The term *objectives* refers to the specific goals of a program—what the program planners intend to achieve. Consider the excerpts from the description of a new health-related graduate-level course given in Example 2.2.

Example 2.2 The Objectives of a New Course

Course Description

The new two-semester course is designed to teach first- and second-year graduate students to conduct evaluations of health programs. Among the

primary aims of the course is the development of a handbook on evaluation that teaches students the basic principles of evaluation and offers an annotated bibliography so that readers can obtain more information when they need it. At the end of the two semesters, each student will be expected to plan the evaluation of a program. The plan is to include evaluation questions, standards, study design and sampling methods, and data collection measures.

Based on this excerpt, the objectives of this course are as follows:

- For the curriculum developer:
  To produce an evaluation handbook with an annotated bibliography
- For the student:
  To prepare an evaluation plan that includes questions, standards, research design, sampling methods, and data collection measures

---

Objectives can involve any of the users or participants in the evaluation: patients, students, health care practitioners, the health care system, and so on. The evaluation questions for the health program evaluation course described in Example 2.2 might include the following:

1. Was a health program evaluation handbook produced?
2. Did each student prepare an evaluation plan with questions, standards, study design, sampling, and data collection measures?

The identification of these two questions immediately raises some additional questions: By when should the handbook be produced? How will we determine if it is any good? What are the characteristics of a satisfactory evaluation plan, and who will judge the students' plans? These questions must be answered in subsequent evaluation activities. In the next step of the evaluation, for example, we will consider ways of setting standards for determining achievement of objectives as well as program effectiveness and efficiency.

When identifying evaluation questions based on goals and objectives, evaluators must be certain that they have identified all of the important goals and objectives, that the evaluation questions cover all of the important objectives, and that all of the questions can be answered with the resources available.

Participants and Effectiveness

In health program evaluation, evaluation questions often aim to describe the demographic and health characteristics of participants in a program and to link effective outcomes to specific participants. An evaluator might be asked, for example, to find out whether a diabetes education program was effective for all

patients or only for a portion—say, patients under 18 years of age. Returning to the new health program evaluation course discussed above, consider the questions about the program's participants shown in Example 2.3.

Example 2.3 Evaluation Questions, Participants, and Program Effectiveness

The developer of the new health program evaluation course for first- and second-year graduate students was concerned with finding out whether the program was effective for all types of students. One measure of effectiveness for a student who has completed the course is the student's ability to prepare a satisfactory evaluation plan. The evaluator asked the following evaluation questions:

- What are the demographic characteristics of each year's students?
- Is the program equally effective for differing students (for example, males and females)?
- Do first- and second-year students differ in their learning?
- At the end of their second year, do the current first-year students maintain their learning?

As noted previously, evaluation questions should be answerable with the resources available. Suppose that the evaluation described in Example 2.3 is only a one-year study. In that case, the evaluator cannot answer the question regarding whether this year's first-year students maintained their learning over the next year. Practical considerations often temper the ambitions of an evaluation.

Program Activities, Organization, and Effectiveness

Evaluators often find that learning about a program's specific activities and its organization is important to their understanding of its success or failure and whether it is applicable to other settings. The following are some typical questions evaluators ask when focusing on program activities:

- What were the key activities?
- To what extent were the activities implemented as planned?
- How well was the program administered?
- Did the program's influence carry over to other programs, institutions, or consumers?
- Was the effectiveness of the program influenced by changes in the social, political, or financial circumstances under which it was conducted?

Consider the case study in Example 2.4, in which specific questions are posed about program activities and organization.

Example 2.4 Evaluation Questions About Program Activities and Organization

A nine-member panel of experts in public health, nursing, health services research, and evaluation met to define the kinds of learning that are appropriate for a course in health program evaluation. The course's evaluation is to take place over a 4-year period so as to enlist two groups of first- and second-year students. Several of the graduate school's best instructors were selected to help design the curriculum and the handbook and to teach the course. The evaluator asks:

- To what extent is the selection of the best teachers responsible for the quality of student learning and of the handbook?
- Does the new course affect students' subsequent education activities?
- Over the 4-year period of the evaluation, do any changes occur in the school's support for the program or the number and types of faculty members who were willing to participate?

Economics and Costs

Program evaluations can be designed to answer questions about the resources that are consumed to produce program outcomes. The resources used, or the program costs, include any expenditures, whether in the form of money, personnel, time, and facilities (e.g., office equipment and buildings). The outcomes may be monetary (e.g., numbers of dollars saved) or substantive (e.g., years of life saved). When questions focus on the relationship between costs and monetary outcomes, the evaluation is termed a *cost-benefit analysis.* When questions are asked about the relationship between costs and substantive outcomes, the evaluation is called a *cost-effectiveness analysis.* The distinction between evaluations concerned with cost-effectiveness and those addressing cost-benefit is illustrated by these two examples:

- *Cost-effectiveness evaluation:* What are the comparative costs of Programs A and B in providing the means for pregnant women to obtain prenatal care during the first trimester?
- *Cost-benefit evaluation:* For every $100 spent on prenatal care, how much is saved on neonatal intensive care?

In the past, health program evaluators have asked questions about costs relatively infrequently, for a number of reasons. Among these are the difficulties inherent in defining costs and measuring benefits in the area of health care. Also, adding the complexities of an economic analysis to an already complex evaluation design may not be a good idea. After all, why study the costs of an intervention of (as yet) unproven effectiveness? To conduct cost studies, evaluators require knowledge of economics and statistics. It is often wise to include a health economist on the evaluation team if you plan to analyze costs.

Example 2.5 illustrates the types of questions that program evaluators pose about the costs, effects, benefits, and efficiency of health care programs.

Example 2.5 Evaluation Questions: Costs

- What is the relationship between the cost and the effectiveness of three prenatal clinic staffing models: physician based, mixed staffing, and clinical nurse specialists with physicians available for consultation? Costs include number of personnel, hourly wages, number of prenatal appointments made and kept, and number of hours spent delivering prenatal care. Outcomes (effectiveness) include maternal health (such as complications at the time of delivery), neonatal health (such as birth weight), and patient satisfaction.
- How efficient are a health care center's ambulatory clinics? Efficiency is defined as the relationship between the use of practitioner time and the size of a clinic, waiting times for appointments, time spent by faculty in the clinic, and time spent supervising house staff.
- How do the most profitable private medical practices differ from the least profitable in terms of types of ownership, collection rates, no-show rates, percentage of patients without insurance coverage, charge for a typical follow-up visit, space occupancy rates, and practitioner costs?
- To what extent does each of three programs to control hypertension produce an annual savings in reduced health care claims that is greater than the annual cost of operating the program? The benefits are costs per hypertensive client (the costs of operating the program in each year, divided by the number of hypertensive employees being monitored and counseled that year). Because estimates of program costs are produced over a given 2-year period but estimates of savings are produced in a different (later) period, benefits have to be adjusted to a standard year. To do this, one must adjust the total claims paid in each calendar year by the consumer price index for medical care costs to a standard 2003 dollar. The costs of operating the programs are similarly adjusted to 2003 dollars, using the same index.

As these questions illustrate, evaluators must define costs and effectiveness or benefits and, when appropriate, must describe the value of the monetary costs. Evaluators who answer questions about the costs of health programs sometimes perform a "sensitivity analysis" when measures are not precise or the estimates are uncertain. For example, in a study of the comparative cost-effectiveness of two state-funded school-based health care programs, the evaluators might analyze the influence of increasing each program's funding first by 5% and then by 10% to test the "sensitivity" of the program's effectiveness to changes in funding level. Through this analysis, the evaluators will be able to tell whether or not increases in measures of effectiveness keep pace with increases in costs.

Program Environment

All programs take place in particular institutional, social, and political environments. For instance, Program A, which aims to improve the preventive health care practices of children under age 14, takes place in rural schools and is funded by the federal government and the state. Program B has the same aim, but it takes place in a large city and is supported by the city and a private foundation.

When an evaluation takes place over several years (say, 3 years or longer), the social and/or political environment in which the program exists can change. New people and policies may emerge, and these may influence the program and the evaluation. Among the environmental changes that have affected programs in health care are alterations in reimbursement policies for hospitals and physicians, the development of new technologies, and advances in medical science. For example, the decrease in the infant mortality rate seen in the United States in recent decades is generally conceded to be the result of programs in prenatal care as well as increases in Medicaid spending for prenatal care, medical advances in treating the underdeveloped lungs of premature infants in their first hours of life, and other improvements in neonatal intensive care.

When evaluators are investigating a program's environment, they will often consider the program's setting and funding as well, as illustrated in Figure 2.1. In addition to asking questions about a program's setting and funding, questions about program management and politics are also relevant:

- *The managerial structure:* Who is responsible for the program's outcomes? How effective is the managerial structure? If the individuals or groups who are running the program were to leave, would the program continue to be effective?
- *The political context:* Is the political environment (meaning within and/or outside the institution) supportive of the success of the program? Is the program well funded?

Program/Intervention Settings

Type of setting(s):
(check all that apply)

[ ] Community hospital clinic
[ ] Community freestanding clinic
[ ] Community physicians' office
[ ] Academic hospital clinic
[ ] Residential treatment facility
[ ] Private residence
[ ] Other facility type not shown above, specify:

_____

_____

Geographic location(s):

A. Country:
[ ] U.S.
[ ] European
[ ] Other, specify:
_____
_____

B. State(s):
1. _____
2. _____ *(enter up to*
3. _____ *five different*
4. _____ *state codes or*
5. _____ *abbreviations)*

[ ] CHECK HERE IF STUDY USED >5 STATES

C. Local (e.g., county/city)
1. _____
2. _____
3. _____
4. _____
5. _____

[ ] CHECK HERE IF STUDY USED >5 CITIES/COUNTIES

Funding source(s):
(check all that apply)

[ ] Federal government, specify: _____
[ ] State government, specify: _____
[ ] Local government (county/city), specify: _____
[ ] Private foundation, specify: _____
[ ] Other, specify: _____
[ ] None stated

Figure 2.1. A Form to Use in Surveying the Program's Environment

Setting Standards of Effectiveness: What They Are and How to State Them

Program evaluations aim to provide convincing evidence of programs' effectiveness. Evaluators measure program effectiveness against particular standards, or specific criteria. Consider the following evaluation questions and their associated standards:

- *Evaluation question:* Did students learn to formulate evaluation questions?
- *Standard:* Of all students in the new program, 90% will learn to formulate evaluation questions. Learning to formulate questions means identifying

and justifying program goals, objectives, and benefits and stating the questions in a comprehensible manner. Evidence that the questions are comprehensible will come from review by at least three potential users of the evaluation.

And:

- *Evaluation question:* Did the current group of first-year students maintain their learning by the end of their second year?
- *Standard:* No decreases in learning will be found between students' second and first years.

In this case, unless 90% of students learn to formulate questions by the end of the first year *and* first-year students maintain their learning over time, the evaluator cannot say the program is effective.

The standards are the key to the evaluation's credibility. The more specific they are, the easier they are to measure. To get at specificity, evaluators must clearly define all potentially ambiguous terms in the evaluation questions and standards. Ambiguity arises when uniformly accepted definitions or levels of performance are unavailable. For example, in the evaluation question "Has the Obstetrical Access and Utilization Initiative improved access to prenatal care for high-risk women?" the terms *improved access to prenatal care* and *high-risk women* are potentially ambiguous. To clarify these terms and thus eliminate ambiguity, the evaluators might find it helpful to engage in a dialogue like the one presented in Example 2.6.

Example 2.6 Clarifying Terms and Setting Standards: A Dialogue Between Evaluators

Evaluator 1: "Improved" means bettered or corrected.

Evaluator 2: For how many women and over what duration of time must care be bettered? Will all women be included? Or 90% of all women, but 100% of teens?

Evaluator 1: "Improved access" means more available and convenient care.

Evaluator 2: What might render care more available and convenient? Care can be made more available and convenient if some or all the following occur: changes in the health care system to include the provision of services relatively close to clients' homes; shorter waiting times at clinics; for some women, financial help, assistance with transportation to care, and aid with child care; and education regarding the benefits of prenatal care and compliance with nutrition advice.

Evaluator 1: "High-risk women" are women whose health and birth outcomes have a higher-than-average chance of being poor.

Evaluator 2: Which, if not all, of the following women will you include? Teens? Users of drugs or alcohol? Smokers? Low-income women? Women with health problems such as gestational diabetes or hypertension?

After they have clarified the terms of the question "Has the Obstetrical Access and Utilization Initiative improved access to prenatal care for high-risk women?" the evaluators might develop standards such as those listed in Example 2.7.

Example 2.7 Standards for Access to and Use of Prenatal Care Services

- At least four classes in nutrition and "how to be a parent" will be implemented, especially for teenagers.
- All clinics will provide translation assistance in English, Spanish, Hmong, and Vietnamese.
- Over a 5-year period, 80% of all pregnant women without transportation to clinics and community health centers will receive it.

Notice that these three standards refer to changes in the structure of health care provision: specially designed education, translation assistance, and transportation. A useful way to think about standards is to decide whether you want to measure the program's effectiveness in terms of the structure, process, or outcomes of health care. Health program evaluators often conceptualize the quality of care in these terms.

The *structure of care* refers to the environment in which health care is given as well as the characteristics of the health care practitioners (including the number of practitioners and their educational and demographic backgrounds), the setting (a fee-for-service program or managed care, for example), and the organization of care (for example, how departments and teams are run).

The *process of care* refers to what is done to and for patients and includes the technical and humanistic aspects of care. The processes of care include the procedures and tests used by the health care team in prevention, diagnosis, treatment, and rehabilitation.

The *outcomes of care* are the results, or the consequences for the patient, of the health care systems, settings, and processes. These include measures of morbidity and mortality; social, psychological, and physical functioning; satisfaction with care; and quality of life.

Example 2.8 presents illustrative standards for the evaluation question "Has the Obstetrical Access and Utilization Initiative improved access to care for high-risk women?"

Example 2.8 Structure, Process, and Outcome Standards for an Evaluation Question About Access to Prenatal Care

- *Structure standard:* All waiting rooms will have special play areas for patients' children.
- *Process standard:* All physicians will justify and apply the guidelines prepared by the College of Obstetrics and Gynecology for the number and timing of prenatal care visits to all women.
- *Outcome standard:* Significantly fewer low-birth-weight babies will be born in the experimental group than in the control group, and the difference favoring the experimental group will be at least as large as the most recent findings reported in the literature.

Standards must be purposeful; arbitrary standards may doom a program. Suppose the Obstetrical Access and Utilization Initiative aims to reduce the no-show rate for obstetrical clinic appointments from 30% to 20%. If the rate actually decreases by only 5% rather than 10%, will the program be considered a failure? What about if the rate decreases by 7%? The evaluators should justify their answers to these questions by using data from other evaluations, the views and ideals of experts, and statistical comparisons.

In addition to being meaningful, standards of effectiveness should be realistic and measurable. Consider the following evaluation question and standard:

- *Evaluation question:* How do students in our medical center compare with students in other medical centers in their knowledge of the content of appropriate prenatal care for high-risk women?
- *Standard:* A statistically significant difference in knowledge will be obtained that favors our institution.

Unless the evaluators have access to students in other medical centers and can test or observe them in the time allotted for the evaluation, this standard, although perhaps desirable, is unrealistic and cannot be used.

How to Set Standards

To establish standards, program evaluators often review other programs, rely on the consensus of experts regarding what is clinically meaningful, use data from community-based data sets, and analyze the research literature.

Setting Standards Using Comparisons With Other Programs

Evaluators can make comparisons using one group's performance over time, several groups' performance at one time, or several groups' performance over time. It is sometimes difficult, however, to constitute similar comparison groups, engage their cooperation, and also have enough time to collect data to measure and observe meaningful differences (assuming they exist). It is extremely important to note that, just because an evaluation seems to find that one group is different from another and the difference favors the new program, this does not automatically mean that an effective program is at work. At least two questions must be asked before any such judgment is possible:

1. Were the groups comparable to begin with? (After all, by coincidence, the individuals in one group might be smarter, healthier, more cooperative, or otherwise different from those in another.)

2. Is the magnitude of the difference large enough to be meaningful? With very large samples, small differences (in scores on a standardized test of achievement, for example) can be statistically, but not clinically, significant. (That is, when the evaluation data are analyzed using standard statistical methods, the groups are found not to be the same on some important health-related outcome, but this statistically significant difference means little at the clinical level.)

Consider an evaluator who is asked to study the effectiveness of an 8-week cognitive-behavioral therapy program for children with measurable symptoms of depression. The evaluation design consists of an *experimental group* of children who receive the program and a *control group* of children who do not. In such a study design, the participants in the control group may get an alternative program, may get no program, or may continue doing what they have been doing all along ("usual care"). To guide the evaluation design, the evaluator hypothesizes that the children who make up the two groups are the *same* in terms of their symptoms before and after the program.

The evaluator administers standardized measures of depression symptoms to all the children in both groups before the experimental program begins and within 1 week of its conclusion. After analyzing the data using traditional statistical tests, the evaluator finds that in fact the children in the experimental group improve (have better scores on the depression symptom measure) whereas those in the control group do not. Using these statistical findings, the evaluator rejects the hypothesis that the two groups are the same after the program and concludes that because they differ statistically, the program is effective.

Some of the participant children's teachers, however, challenge the evaluator's conclusion by asking if the statistical difference is clinically meaningful. These teachers are not convinced that the improvement in scores in the experimental

group means much. After all, the depression symptoms measure is not perfect, and the gains indicated may disappear over time. Through this experience, the evaluator learns that if you rely solely on statistical significance as a standard of effectiveness, you may be challenged to prove that the statistics mean something practical in clinical terms.

The difference between statistical and clinical significance is particularly important in program evaluations that focus on impact, outcomes, and costs. Suppose that students' scores on a standardized test of achievement increase from 150 to 160. Does this 10-point increase mean that students are actually more capable? How much money are policy makers willing to allocate for schools to improve scores by 10 points? On the other hand, consider a group of people in which each loses 10 pounds after being on a diet for 6 months. Depending on where each person started (e.g., some may need to lose no more than 10 pounds, some may need to lose at least 50), a loss of 10 pounds may be more or less clinically significant.

Another way to think of standards of clinical significance is through the concept of *effect size.* Consider this conversation between two evaluators:

Evaluator A: We have been asked to evaluate a Web-based program for high school students that aims to decrease their health-related risks through interactive education and the use of online support groups. We particularly want students to stop smoking. How will we know if the program is effective? Has anyone done a study like this?

Evaluator B: I don't know of anyone offhand, but we can e-mail some people I know who have worked with online programs to reduce health risks. Maybe they have some data we can use. Also, we can do a search of the literature.

Evaluator A: What would we look for?

Evaluator B: We want to know if a program like ours has ever been proven to result in significantly fewer smokers. If we find that any programs have had such results, we will then want to know if a sufficiently large number of students have quit smoking so that we have some truly meaningful results. We may need to work with the school district to find out how large the number of students who quit should be. Alternatively, we can use standard statistical methods to decide on the number.

Evaluator B is getting at the concept of effect size when she talks about having a sufficiently large number of students who quit smoking so that the results, if significant, are also meaningful.

How do you determine the effect size? As Evaluator B suggests, the best sources of information on how other programs have performed are consultations with

colleagues and reviews of the literature. But you may not find performance data through such sources, especially if the program you are evaluating is especially innovative. Further, data are not always available because not everything is published. If you cannot find information on other programs similar to yours, you may have to conduct a small-scale or *pilot study* to get estimates of effect sizes for your program. You can also use statistical formulas to calculate effect sizes. Regardless of how you decide on the extent of the effect, you will probably want to ask experts to confirm its practical and clinical significance.

Setting Standards Using Experts

Experts can assist evaluators in setting standards and in confirming the practical or clinical significance of the findings. In this context, an *expert* is defined as any individual or representative of a professional, consumer, or community group who is likely to use the results of an evaluation.

Evaluators use a variety of techniques to consult with and promote agreement among experts. These usually include the selection of representative groups of experts who then take part in structured meetings. For example, for an evaluator who is concerned with setting standards for a program to improve the quality of instruction in health policy and health services research, an appropriate group of advisers would include experts in those fields, experts in education, and consumers of health services research and policy (such as representatives from the public).

The fields of health and medicine make extensive use of expert panels. For example, the National Institutes of Health has used consensus development conferences to help resolve issues related to knowledge about and use of particular medical technologies, such as intraocular lens implantation, as well as the care of patients with specific conditions, such as depression, sleep disorders, traveler's diarrhea, and breast cancer. The American College of Physicians, the American Heart Association, the Institute of Medicine, and the Agency for Health Care Research and Quality are some of the many organizations that consistently bring experts together to establish "guidelines" for practice concerning common problems such as pain, high blood pressure, and depression.

The main purpose of seeking consensus among experts is to define levels of agreement on controversial subjects and unresolved issues. These methods are therefore extremely germane to setting standards against which to judge the effectiveness of new programs when no comparison group data are available. True consensus methods, however, are often difficult to implement, because they typically require extensive reviews of the literature on the topic under discussion as well as highly structured methods.

The use of expert panels has proven to be an effective technique in program evaluation for setting standards of performance, as illustrated in Example 2.9.

Example 2.9 Using Experts to Set Standards

Sixteen U.S. teaching hospitals participated in a 4-year evaluation of a program to improve outpatient care in their group practices. Among the study's major goals were improvements in amount of faculty involvement in the practices, in staff productivity, and in access to care for patients. The evaluators and representatives from each of the hospitals used standards for care set by the Institute of Medicine as a basis for setting standards of program effectiveness before the start of the study. After 2 years, the evaluators presented interim data on performance and brought experts from the 16 hospitals together to come to consensus on standards for the final 2 years of the study. To guide the process, the evaluators prepared a special form, part of which appears in Figure 2.2.

Variable	Current Standard	Definitions	Interim Results	Question	Suggestion	Your Decision
Waiting Time	90% of patients should be seen within 30 minutes	Waiting time is time between scheduled appointment and when first seen by primary provider	70% of patients were seen within 30 minutes	Is 90% reasonable?	90%	__ %
	Compared to national databases, program patients should not have unduly long waiting times		Waiting times = 24.3 minutes. National data = 37.3 minutes for doctor's office or private clinic.	Should national data be used as standard?	Yes	__ Yes __ No

Figure 2.2. Selected Portions of a Form Used in Setting Standards of Effectiveness

It is interesting to note that a subsequent survey of the participants in the standard-setting process discussed in Example 2.9 found that they did not use the interim data to make their choices: No association was found between how well a medical center had previously performed with respect to each of 36 selected standards and the choice of a performance level for the remaining 2 years of the evaluation. Interviews with experts at the medical centers revealed that the standards they selected came from informed estimates of what performance might yet become and from the medical centers' ideals; the experts considered the interim data to be merely suggestive.

Program evaluators use a number of methods when they rely on panels of experts to promote understanding of issues, topics, and standards for evaluation, but the most productive of these depend on a few simple practices, as discussed in the following guidelines:

Guidelines for Expert Panels

1. *The evaluator must clearly specify the evaluation questions.* If the questions are not clearly specified, the experts may help in clarification and in specification. Here are examples:
   Not quite ready for standard setting: Was the program effective with high-risk women?
   More amenable to standard setting: Did the program improve the proportion of low-weight births among low-income women?
   Standard: Significantly fewer low-weight births are found in the experimental versus the control group.

2. *The evaluator should provide data to assist the experts.* These data can be about the participants in the experimental program, the intervention itself, and the costs and benefits of participation. The data can come from the published literature, from ongoing research, or from financial and statistical records. For example, in an evaluation of a program to improve birth weight among infants born to low-income women, experts might make use of information about the extent of the problem in the country. They might also want to know how prevalent low-weight births are among poor women and, if other interventions have been used effectively, what their costs were.

3. *The evaluator should select experts based on their knowledge, their influence, or how they will use the findings.* The number of experts an evaluator chooses is necessarily dependent on the evaluation's resources and the evaluator's skill in coordinating groups. (See Example 2.10 for two illustrations concerning the choice of experts.)

4. *The evaluator should ensure that the panel process is carefully structured and skillfully led.* A major purpose of the expert panel is to come to agreement on the criteria for appraising a program's performance. To facilitate agreement, and to distinguish the panel process from an open-ended committee meeting, the evaluator should prepare an agenda for the panel in advance, along with the other materials noted above (such as literature reviews and other presentations of data). When possible, the evaluator should try to focus the panel on particular tasks, such as reviewing a specific set of data and rating the extent to which those data apply to the current program. For example, the evaluator might give the experts data on past performance (e.g., 10 of 16 hospitals had continuous quality improvement systems for monitoring the quality of inpatient care) and ask them to rate the extent to which that standard should still apply (e.g., on a 5-point scale ranging from *strongly agree* to *strongly disagree*).

Example 2.10 Choosing Experts to Set Evaluation Standards

- The New Dental Clinic aimed to improve patient satisfaction. A meeting was held in which three patient representatives, a nurse, a physician, and a technician defined the "satisfied patient" and decided on how much time to allow the clinic to produce satisfied patients.
- The primary goals of the Adolescent Outreach Program are to teach teens about preventive health care and to make sure that all needed health care services (such as vision screening and immunizations) are provided. A group of teens participated in a teleconference to help the program developers and evaluators decide on the best ways to teach teens and to set standards of learning achievement. Also, physicians, nurses, teachers, and parents participated in a conference to determine the types of services that should be provided and how many teens should receive them.

Setting Standards Using Community Data Sets

Large data sets such as those maintained by the Centers for Disease Control and Prevention (CDC), the National Center for Health Statistics (which can be accessed through the CDC Web site at http://www.cdc.gov), and state and local governments come from scientific surveys of whole populations of various kinds and contain accurate information on individual and collective health status as well as on the composition, organization, and financing of the health system. Researchers also maintain data sets, and many make their data available to other investigators.

The information from large data sets is often presented in summary or report form. Such data sets provide benchmarks against which evaluators can measure the effectiveness of new programs. For instance, an evaluator of a hypothetical drivers' education program might say something like this: "I used the county's Surveillance Data Set to find out about use of seat belts. The results show that about 1 in 10 drivers between 18 and 21 years of age in this county do not use seat belts. An effective driver education program should be able reduce that number to 1 in 11 within 5 years."

Suppose you have been asked to evaluate a new program to prevent low-weight births in your state. If you know the current percentage of low-weight births in the state, then you can use that figure as a gauge for evaluating the effectiveness of a new program that aims to lower the rate. Example 2.11 shows some ways in which evaluators use existing data to set standards in program evaluation.

Example 2.11 Using Community Data to Set Standards in Program Evaluation

- The Obstetrical Access and Utilization Initiative serves high-risk women and aims to reduce the numbers of births of babies weighing less than 2,500 grams (5.5 pounds). One evaluation question asks, "Is the birth of low-weight babies prevented?" In the state, 6.1% of babies are low birth weight, but this percentage includes babies born to women who are considered to be at low or medium risk. The standard used as evidence that low-weight births are prevented is as follows: "No more than 6.1% of babies will be born weighing less than 5.5 pounds."
- The city's governing council has decided that the schools should become partners with the community's health care clinics in developing and evaluating a program to reduce motor vehicle crashes among children and young adults ages 10 to 24 years. According to the Centers for Disease Control's findings from the Youth Risk Behavior Surveillance System (accessed through the CDC Web site, http://www.cdc.gov), the leading cause of death (31% of all deaths) among youth of this age is motor vehicle accidents. Council members, community clinic representatives, teachers and administrators from the schools, and young people meet to discuss standards for program effectiveness. They agree that they would like to see a statistically and clinically meaningful reduction in deaths due to motor vehicle crashes over the program's 5-year trial period. They will use the 31% figure as the baseline against which to evaluate reduction.

When you use data from large sets of data as the standards against which you are evaluating a local program, you must make certain that they are truly applicable. The only data available to you may have been collected a long time ago or under circumstances that are very different from those surrounding the program you are evaluating, and so they may simply not apply. For example, data gleaned from an evaluation conducted with men may not apply to women, and data on older men may not apply to younger men. Data from an evaluation conducted with hospitalized patients may not apply to people in the community. Also, there may be geographic and cultural variations in how people perceive and deal with health; standards for the entire community may not always apply locally.

Setting Standards Using the Literature

The literature consists of all published and unpublished reports of evaluation studies. To identify standards for your program, you should use only the most scientifically rigorous evaluations of other programs. You must also be careful to base your

standards on programs and participants reported in the literature that are sufficiently similar to the program and participants you are evaluating. Example 2.12 illustrates how evaluators might use the literature in setting standards in program evaluations.

Example 2.12 Using the Literature to Set Evaluation Standards

The Community Cancer Center has inaugurated a new program to help families deal with the depressive symptoms that often accompany a diagnosis of cancer in a loved one. A main program objective is to reduce symptoms of depression among participating family members.

The evaluators want to convene a group of potential program participants to assist in developing standards of program effectiveness. In specific, the evaluators want assistance in defining "reduction in symptoms." They discover, however, that it is very difficult to find a time when all of the potential participants are available to meet. Also, the center does not have the funds to sponsor a face-to-face meeting. Because of these constraints, the evaluators decide against convening a meeting and instead turn to the literature.

The evaluators go online to find articles that describe the effectiveness of programs to reduce depressive symptoms in cancer patients. Although they find five published articles, only one of the programs evaluated is exactly the same as the Community Cancer Center's program, although it takes place in an academic cancer center. Nevertheless, given the quality of the evaluation and the similarities between the two programs, the evaluators believe that they can apply this other program's standards to the present program. This is what the evaluators found in the article:

> At the 6-month assessment, family members in the first group had significantly lower self-reported symptoms of depression on the Depression Scale than did family members in the second group (8.9 versus 15.5). The mean difference between groups adjusted for baseline scores was –7.0 (95% confidence interval, –10.8 to –3.2), an effect size of 1.08 standard deviations. These results suggest that 86% of students who underwent the program reported lower scores of depressive symptoms at 6 months than would have been expected if they had not undergone the program.

The evaluators decide to use the same measure of depressive symptoms (the Depression Scale) as did the evaluator of the published study and to use the same statistical test to determine the significance of the results.

When adopting standards from the literature, you must compare the characteristics of the program you are evaluating and the program or programs whose evaluation standards you plan to adopt. You need to make certain that the participants, settings, interventions, and main outcome variables are similar, if not identical.

Then, when you do your evaluation, you must choose exactly the same measures or instruments and statistical methods to interpret the findings.

Evaluation Standards and Economic Evaluations

In a cost-effectiveness analysis, a program is considered to be effective if no other program is available at lower cost (that is, other programs may be available, but they cost more). Four generic standards apply in economic evaluations:

- *Cost-effectiveness evaluation:* Program A is effective and is the lowest-cost program around.
- *Cost-benefit analysis:* Program A has merit if its benefits are equal to or exceed its costs; the benefit-to-cost ratio of Program A is equal to or greater than 1.0 and exceeds the benefit-to-cost ratio of Program B.
- *Cost minimization analysis:* Programs A and B have identical benefits, but Program A has lower costs.
- *Cost utility analysis:* Program A produces *N* (the evaluation figures out exactly how many) quality adjusted life years at lower cost than Program B.

Example 2.13 illustrates the uses of economic standards in evaluations.

Example 2.13 Setting Standards for Economic Evaluations

RISK-FREE is a new preventive health education program. The evaluators have three study aims and associated hypotheses. Two of the study aims (Aims 2 and 3) pertain to an economic evaluation.

- *Aim 1:* To evaluate the comparative effectiveness of a patient and physician educational intervention to prevent risky health behaviors in adults relative to usual care
  Hypothesis 1: When baseline levels are controlled for, the experimental patients will have a significantly lower probability of risky health behaviors over a 12-month period than patients in the usual care arm.
  Hypothesis 2: When baseline levels are controlled for, the experimental patients will have significantly better health-related quality of life over a 12-month period than patients in the usual care arm.

As part of an impact evaluation, the evaluators will also examine proximal outcomes (mediating factors in the relationship between the randomized intervention and the effectiveness end points), testing two additional hypotheses:

> *Hypothesis 3:* When baseline levels are controlled for, the experimental patients will demonstrate significantly greater self-efficacy and knowledge over a 12-month period than patients in the usual care arm.

Hypothesis 4: When baseline levels are controlled for, the experimental providers will demonstrate significantly greater knowledge and more positive attitudes, and patients will report that their providers engage in more counseling, over a 12-month period than providers in the usual care arm.

- *Aim 2:* To evaluate the comparative costs of the patient and physician educational intervention relative to usual care
  Hypothesis 5: When baseline levels are controlled for, the experimental patients will have significantly lower utilization and net (intervention + nonintervention) costs over a 12-month period than patients in the usual care arm. (This hypothesis is based on the expectation that the intervention will increase nonintervention costs by more than the cost of the intervention itself, leading to net cost savings.)
- *Aim 3:* To evaluate the cost-effectiveness of the patient and physician educational intervention relative to usual care
  Hypothesis 6: The experimental intervention will be cost-effective relative to care as usual, based on generally accepted threshold values for incremental cost-effectiveness ratios (dollars per quality-adjusted life year).

If Aim 2 demonstrates that the intervention is cost saving with equal outcomes or cost neutral with better outcomes, then the intervention is cost-effective by definition and the Aim 3 analyses are unnecessary.

Evaluation Questions and Standards: Establishing a Healthy Relationship

An evaluation question may have just one standard, or several standards may be associated with it. For example:

Question: Did nurses learn to abstract medical records reliably?

Standard 1: 80% of all nurses learn to abstract medical records reliably.

Standard 2: A statistically significant difference in learning is observed between nurses at Medical Center A and those at Medical Center B. Nurses at Medical Center A have participated in a new program, and the difference is in their favor.

The reason the question in this example has two standards is that if 80% of nurses at Medical Center A are found to be able to abstract medical records reliably, we cannot really attribute this positive result to a program unless we have access to the abstractions of nurses who were not in a program. After all, nearly all nurses might know from the start how to abstract medical records. So why not just rely on the second standard? Because differences alone, no matter how great, may not be enough.

Suppose that the nurses' learning at Medical Center A was actually significantly higher than the learning that took place at Medical Center B. If the learning levels

of both groups were low (say, average scores of 50% in the new program group and 20% in the comparison), then the difference, although statistically meaningful, might not be educationally or clinically meaningful, and thus it has little practical merit. We can conclude, for example, that, even though the program successfully elevated performance, the amount of improvement was insufficient, and so the program was simply not satisfactory.

When to Set Standards

The evaluator should have standards in place before continuing with the evaluation's design and analysis. Consider the following two examples.

Example 1

Program goal: To teach nurses to abstract medical records reliably

Evaluation question: Have nurses learned to abstract medical records reliably?

Standard: 90% of all nurses learn to abstract medical records reliably

Program effects on: Nurses

Effects measured by: Reliable abstraction

Design: A survey of nurses' abstractions

Data collection: A test of nurses' ability to abstract medical records

Statistical analysis: Computation of the percentage of nurses who abstract medical records reliably

Example 2

Program goal: To teach nurses to abstract medical records reliably

Evaluation question: Have nurses learned to abstract medical records reliably?

Standard: A statistically significant difference in learning is observed between nurses at Medical Center A and nurses at Medical Center B. Nurses at Medical Center A have participated in a new program, and the difference is in their favor.

Program effects on: Nurses at Medical Center A

Effects measured by: Reliable abstraction

Design: A comparison of two groups of nurses

Data collection: A test of nurses' ability to abstract medical records

Statistical analysis: A *t* test to compare average abstraction scores between nurses at Medical Center A and nurses at Medical Center B

The evaluation questions and standards contain the independent and dependent variables on which the evaluation's design, measurement, and analysis are subsequently based. Independent variables are sometimes called *explanatory* or *predictor* variables because, as these variables are present before the start of the program (that is, are independent of it), evaluators use them to "explain" or "predict" outcomes. In the example above, reliable abstraction of medical records (the outcome) is to be explained by nurses' participation in a new program (the independent variable). In evaluations, the independent variables often are the program (experimental and control), demographic features of the participants (such as gender, income, education, experience), and health characteristics of the participants (such as functional status and physical, mental, and social health).

Dependent variables, also termed *outcome* variables, are the factors the evaluator expects to measure. In program evaluations, these include health status, functional status, knowledge, skills, attitudes, behaviors, costs, and efficiency.

Thus the evaluation questions and standards necessarily contain the independent and dependent variables: those on whom the program is to have effects and measures of those effects, as illustrated in Example 2.14.

Example 2.14 Questions, Standards, and Independent and Dependent Variables

Program goal: To teach nurses to abstract medical records reliably

Evaluation question: Have nurses learned to abstract medical records reliably?

Standard: A statistically significant difference in learning is observed between nurses at Medical Center A and nurses at Medical Center B. Nurses at Medical Center A have participated in a new program, and the difference is in their favor.

Program effects explained by (independent variable): Participation versus no participation in a new program

Effects measured by this outcome (dependent variable): Reliable abstraction

---

The QSV Report: Questions, Standards, Variables

The relationships among evaluation questions, standards, and variables can be depicted in a reporting form such as the one shown in Figure 2.3. You will find that reporting questions, standards, and variables in this format is useful as you go about planning your evaluation and accounting for its methods. As the figure shows, the evaluation questions appear in the first column of the QSV (questions, standards, variables) report form, followed in subsequent columns by the standards associated with each question, the independent variables, and the dependent variables.

The QSV report in Figure 2.3 shows information on an evaluation of an 18-month program combining diet and exercise to improve health status and quality of life for persons 75 years of age or older who are living at home. Participants will be randomly assigned to the experimental or control groups according to the streets on which they live. Participants in the evaluation who need medical services can choose one of two clinics offering differing models of care delivery, one that is primarily staffed by physicians and one that is primarily staffed by nurses. The evaluators will be investigating whether any differences exist between male and female participants after program participation and the role of patient mix in those differences. (*Patient mix* refers to those characteristics of patients that might affect outcomes; these include sociodemographic characteristics, functional status scores, and presence of chronic disorders such as diabetes and hypertension.) The evaluators will also be analyzing the cost-effectiveness of the two models of health care delivery.

Evaluation Questions	**Standards**	**Independent or Explanatory Variables**	**Dependent or Outcome Variables**
To what extent has quality of life improved?	A statistically and practically significant improvement in quality of life over a 1-year period A statistically and practically significant improvement in quality of life between participants and nonparticipants	Gender, group participation (experimental and control participants), patient mix (sociodemographic characteristics, functional status scores, presence or absence of chronic disorders such as diabetes and hypertension)	Quality of life includes social contacts and support, financial support, perceptions of well-being
To what extent has health status improved?	A statistically and practically significant improvement in quality of life over a 1-year period A statistically and practically significant improvement in quality of life between participants and nonparticipants	Gender, group participation (experimental and control participants), patient mix (sociodemographic characteristics, functional status scores, presence or absence of chronic disorders such as diabetes and hypertension)	Health status includes functional status and perceptions of general health and physical functioning; measures of complications from illness (for diabetes would include cardiac, renal, ophthalmologic, or foot; for hypertension would include blood pressure control)
What is the relationship between cost and effectiveness of two clinic staffing models: primarily physicians and primarily nurses?	Effectiveness will be demonstrated by lower cost per visit and satisfactory health status and quality of life	Two models of care (primarily physician-based and primarily nurse-based)	Quality of life, health status, costs of personnel, hours delivering care, number of appointments made and kept

Figure 2.3. The QSV Reporting Form

As you can see, the QSV reporting form extracts the questions, standards, and variables from the description of the evaluation of the diet and exercise program for persons over 75 years of age. This evaluation has three questions: one about the program's influence on quality of life, one about the program's influence on health status, and one about the cost-effectiveness of two methods for staffing clinics. Each of the three questions has one or more standards associated with it. The independent variables for the questions about quality of life and health status are gender, group participation, and patient mix, and each of these terms is explained. The dependent variables are also explained in the QSV report. For example, the report notes that "quality of life" includes social contacts and support, financial support, and perceptions of well-being.

Summary and Transition to the Next Chapter on Evaluation Research Design

A program evaluation is conducted to determine whether a given program is meritorious. Is it worth the costs, or might a more efficient program accomplish even more? Evaluation questions focus the evaluation. They may be aimed at many different areas, such as the program's environment; the extent to which program goals and objectives have been met; the degree, duration, and distribution of benefits and effects; and the implementation and effectiveness of different program activities and management strategies.

Program evaluations are concerned with providing convincing evidence of programs' effectiveness. The standards are the specific criteria against which effectiveness is measured. The evaluator should set the questions and standards in advance of any evaluation activities because they prescribe the evaluation's design, data collection, and analysis. One question may have more than one standard associated with it. Evaluators set standards based on statistical comparisons, the opinions of experts, and reviews of the literature, past performance, and community data sets.

The next chapter will tell you how to design an evaluation so that you will be able to link any changes found in health, health practices, education, and attitudes to an experimental program and not to other competing events. For example, suppose you are evaluating a health education program that aims to encourage men and women over the age of 50 to get a colonoscopy—a rather uncomfortable procedure that has been shown to protect against colon cancer. You might erroneously conclude that your program is effective if you observe a significant increase in colonoscopies among program participants unless your evaluation's design is sufficiently sophisticated to distinguish between the effects of the health education program and those of other sources of education, such as television and newspapers. The next chapter discusses the most commonly used evaluation research designs.

Exercises

EXERCISE 1

Directions

Read the example below and, using only the information offered, list the evaluation questions.

> The University Medical Center is concerned with continuously auditing its transfusion practices to ensure the safety of the blood supply. Accordingly, the Committee on Blood Derivatives has been formed to establish guidelines for transfusing red blood cells, fresh-frozen plasma, platelets, and cryoprecipitated AHF. An education program is offered to all interested physicians and nurses to teach them about the guidelines, assist in ensuring appropriate transfusion practices, and, in general, improve the quality of care at the institution. A 2-year evaluation is conducted. Among the medical center's concerns is that all physicians use the guidelines.

EXERCISE 2

Directions

Read the example below and state the evaluation questions and associated standards as well as the independent and dependent variables.

> The director of the Infectious Disease Center wants all of the center's nonphysician staff members to acquire knowledge regarding some of the ethical issues pertaining to the care of patients with infectious diseases, including hepatitis and tuberculosis. These specifically include issues pertaining to patient privacy. The director is working with the center's Bioethics Committee and members of its Division of Medical Education to develop a program for nonphysicians. The plan is to institute the program and monitor its effects on staff each year for 5 years.

Suggested Readings

To learn more about questions and standards in economic evaluations, see the following evaluation reports, all of which are available in their entirety online through MEDLINE. (For instructions on how to find MEDLINE, see the "Suggested Readings" section in Chapter 1.)

Fleming, M. F., Mundt, M. P., French, M. T., Manwell, L. B., Stauffacher, E. A., & Barry, K. L. (2002). Brief physician advice for problem drinkers: Long-term efficacy and benefit-cost analysis. *Alcoholism: Clinical and Experimental Research, 26,* 36–43.

Objective: To describe the 48-month efficacy and benefit-cost analysis of Project TrEAT (Trial for Early Alcohol Treatment), a randomized controlled trial of brief physician advice for the treatment of problem drinking.

Jha, A. K., Perlin, J. B., Kizer, K. W., & Dudley, R. A. (2003). Effect of the transformation of the Veterans Affairs health care system on the quality of care. *New England Journal of Medicine, 348,* 2218–2227.

Objective: To determine the change in the quality of health care for veterans resulting from a mid-1990s systemwide reengineering of the VA health care system to improve quality of care and to compare the quality of care in the VA system with that of the Medicare fee-for-service program.

Jones, K., Colson, P. W., Holter, M. C., Lin, S., Valencia, E., Susser, E., & Wyatt, R. J. (2003). Cost-effectiveness of critical time intervention to reduce homelessness among persons with mental illness. *Psychiatric Services, 54,* 884–890.

Objective: To investigate the cost-effectiveness of the critical time intervention program, a time-limited adaptation of intensive case management, which has been shown to reduce recurrent homelessness significantly among men with severe mental illness.

Wheeler, J. R. (2003). Can a disease self-management program reduce health care costs? The case of older women with heart disease. *Medical Care, 41,* 706–715.

Objective: To assess the impact of a heart disease management program on use of hospital services, estimate associated hospital cost savings, and compare potential cost savings with the cost of delivering the program.

About Quality of Care

Donabedian, A. (1980). *Explorations in quality assessment and monitoring: Vol. 1. The criteria and standards of quality.* Ann Arbor, MI: Health Administration Press.

Donabedian, A. (1982). *Explorations in quality assessment and monitoring: Vol. 2. The definition of quality and approaches to its assessment.* Ann Arbor, MI: Health Administration Press.

Donabedian, A. (1983). *Explorations in quality assessment and monitoring: Vol. 3. Methods and findings of quality assessment and monitoring: An illustrated analysis.* Ann Arbor, MI: Health Administration Press.

Institute of Medicine. (2000). *Crossing the quality chasm.* Washington, DC: National Academy Press.

See Also

Drummond, M. F., Stoddard, G. L., & Torrance, G. W. (1997). *Methods for economic evaluation of health programs.* New York: Oxford University Press.

This book is an outstanding source of basic information on economic evaluation methods.

Fink, A., Kosecoff, J., & Brook, R. H. (1986). Setting standards of performance for program evaluations: The case of the teaching hospital general medicine group practice program. *Evaluation and Program Planning, 9,* 143–151.

This article describes the methods and usefulness of setting standards in a national study to improve the quality of care and education for health care practitioners. It is one of the few articles to date that discusses methods for standard setting, so even though it was published some time ago, it is still relevant.

Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D., Hearst, N., & Newman, T. B. (2001). *Designing clinical research* (2nd ed.). Philadelphia: Lippincott Williams & Wilkins.

For information on the origins of evaluation questions as well as other types of research questions, see Chapter 2 of this volume.