* ONE *

WHAT IS EVALUATION?

As promised in the preface, this book's approach is to give you a "bare-bones," nuts-and-bolts guide about how to do an **evaluation.'** Although we will not be spending a huge amount of time on evaluation theory, it is certainly a good idea to start with a clear notion of what it is we are getting ourselves into.

BASIC DEFINITIONS

In terms of the evolution of the human race, evaluation is possibly the most important activity that has allowed us to evolve, develop, improve things, and survive in an ever-changing environment. Every time we try something new—a farming method, a manufacturing **process,** a medical treatment, a social change **program,** a new management team, a **policy** or **strategy,** or a new information system—it is important to consider its value. Is it better than what we had before? Is it better than the other options we might have chosen? How else might it be improved to push it to the next level? What did we learn from trying it out?

> Professional evaluation is defined as the **systematic** determination of the **quality** or **value** of something (Scriven, 1991).

Things that we might (and should) evaluate systematically include the following[2]:

- **Projects,** programs, or organizations
- Personnel or performance

- Policies or strategies
- Products or services
- Processes or systems
- Proposals, contract bids, or job applications

There is a fundamental logic and methodology that ties together the evaluation of these different kinds of evaluands. For example, some of the key learnings from the evaluation of **products** and personnel often apply to the evaluation of programs and policies and vice versa. This *transdisciplinary* way of thinking about evaluation provides a constant source of innovative ideas for improving how we evaluate. For this reason, this book contains illustrative examples drawn from a variety of settings and evaluation tasks.

Evaluations are generally conducted for one or two main reasons: to find areas for improvement and/or to generate an **assessment** of *overall* quality or value (usually for reporting or decision-making purposes). Defining the nature of the evaluation question is key to choosing the right methodology.

Some other terms that appear regularly in this book are merit, worth, quality, and value. Scriven (1991) defines these as follows:

Merit is the "intrinsic" value of something; the term is used interchangeably with *quality*.

Worth is the value of something to an individual, an organization, an institution, or a collective; the term is used interchangeably with *value.*

This distinction might seem to be a fine one, but it can come in handy. For example, in the evaluation of products, **services,** and programs, it is important to critically consider the extent to which improvements in *quality* (e.g., adding more "bells and whistles") would actually provide enough incremental *value* for the individuals and/or organization concerned to justify their cost.

More often than not in evaluation, we are looking at whether something is "worth" buying, continuing to fund, enrolling in, or implementing on a broader scale. Accordingly, most "big picture" evaluation questions are questions of value (to recipients/users, funders/taxpayers, and other relevant parties) rather than of pure merit. There are exceptions, however, and that is why I have kept both considerations in play.

FITTING EVALUATION APPROACH TO PURPOSE

For any given evaluation, a range of possible approaches is available to the practitioner and the **client.** The option that is most often discussed in evaluation circles pertains to whether an evaluation should be conducted independently (i.e., by one or more outside contractors) or whether the program or product designers or staff should be heavily involved in the evaluation process.

If the primary purpose of the evaluation is for *accountability*, it is often important to have an **independent evaluation** conducted (i.e., nobody on the evaluation team should have a significant vested interest in whether the results are good or bad). This is not always a requirement (e.g., managers in all kinds of organizations frequently report on the performance of their own units, products, and/or people), but this credibility or independence issue is definitely one to consider when choosing how to handle an accountability-focused evaluation.

There are many cases where independence is not essential, but building **organizational learning capacity** is key; that is, a primary goal is to improve **organizational learning** (i.e., the organization's ability to learn from its successes and failures). In such cases, an evaluation can (and should) be conducted with a degree of **stakeholder** participation. Many high-quality professional evaluations are conducted collaboratively with organizational staff, internal human resources consultants, managers, **customers** or recipients, or a combination of these groups.

A **learning organization** is one that acquires, creates, evaluates, and disseminates knowledge—and uses that knowledge to improve itself—more effectively than do most organizations. The best learning organizations tend to use both independent and **participatory evaluations** to build learning capacity, gather multiple perspectives on how they are doing, and keep themselves honest (Davidson, 2003).

THE STEPS INVOLVED

Whether the evaluation is conducted independently or in a participatory mode, it is important to begin with a clear understanding of what evaluation is and what kinds of evaluation questions need to be answered in a particular case. Next, one needs to identify relevant "values," collect appropriate data, and systematically combine the values with the descriptive data to convey, in a useful and concise way, defensible answers to the key evaluation questions (see Exhibit 1.1).

Exhibit 1.1 Overview of the Book's Step-by-Step Approach to Evaluation

CHAPTER 1 Understanding the basics about evaluation

CHAPTER 2 Defining the main purposes of the evaluation and the "big picture" questions that need answers

CHAPTER 3 Identifying the evaluative **criteria** (using needs assessment and other techniques)

CHAPTER 4 Organizing the list of criteria and choosing sources of evidence (mixed method data)

CHAPTER 5 Dealing with the causation issue: how to tell the difference between **outcomes** or **effects** and coincidental changes not caused by the evaluand

CHAPTER 6 Values in evaluation: understanding which values should legitimately be applied in an evaluation arid how to navigate the different kinds of "subjectivity"

CHAPTER 7 Importance weighting: figuring out which criteria are the most important

CHAPTER 8 Merit determination: figuring out how well your evaluand has done on the criteria (excellent? good? satisfactory? mediocre? unacceptable?)

CHAPTER 9 Synthesis methodology: systematic methods for condensing evaluative findings

CHAPTER 10 Putting it all together: fitting the pieces into the Key Evaluation Checklist framework

CHAPTER 11 Meta-evaluation: how to figure out whether your (or someone else's) evaluation is any good

THE INGREDIENTS OF A GOOD EVALUATION

The overarching framework used for planning and conducting an evaluation and presenting its results is Scriven's (2003) Key Evaluation Checklist (KEC) with a few modifications and simplifications. This is a guiding framework for the evaluation team members (be they organizational members, **external evaluators,** or a mix) to make sure that all important ingredients that will allow valid **evaluative conclusions** to be drawn are included.

The KEC should be thought of both as a checklist of necessary ingredients to include in a solid evaluation and as a framework to help guide evaluation planning and reporting. Because the KEC was designed primarily for application to program evaluation, some of the points might need refraining when the KEC is used for other **evaluands** or **evaluees** (the term used in personnel evaluation). In a posting to a listserv on November 16,2002, Scriven describes how and why the KEC was developed:

> The Key Evaluation Checklist evolved out of the work of a committee set up by the U.S. Office of Education which was to hand out money to disseminate the best educational products to come out of the chain of Federal Labs and R&D Centers (some of which still exist). The submissions were supposed to have supporting evidence, but these documents struck me as frequently making a few similar mistakes (of omission, mostly). I started making a list of the recurring holes, i.e., the missing elements, and finished up with a list of what was needed in a good proof of merit, a list which we used and improved.

A brief overview of the KEC is shown in Exhibit 1.2. Each line of KEC checkpoints represents another layer in the evaluation. We begin with the Preliminaries (Checkpoints I—III), which give us some basic information about the evaluand and the evaluation. From there, we move to the Foundations (Checkpoints 1-5), which provide the basic ingredients we need, that is, descriptive information about the program, who it serves (or should serve), and the values we will apply to evaluate it. The third level, which Scriven called the Sub-evaluations (Checkpoints 6-10), includes all of the explicitly evaluative elements in an evaluation (i.e., where we apply values to descriptive facts to derive evaluative conclusions at the analytical level). Finally, we reach the Conclusions section (Checkpoints 11-15), which includes overall answers to the evaluation questions plus some follow-up elements. **(SEE PAGES 6-7)**

Scriven (1991) asserts that evaluations should generally cover all of these checkpoints (except for Checkpoints 12 and 13, which are optional) to draw valid conclusions. Each point listed in the KEC is backed by a carefully thought-out rationale showing why omission of the particular point is likely to compromise the validity of conclusions. Although this should not be taken to mean that all checkpoints must *always* be included in all evaluations, it does mean that decisions to omit certain elements should be carefully justified. This is particularly important for Checkpoints 5 through 9 and 11, which form the core of the evaluation.

Obviously, there is a lot more to the KEC than one can fit on a one-page summary. Throughout this book, we work through many of the KEC checkpoints, paying particular attention to the truly evaluative* ones (from Checkpoint 5 [Values] through Checkpoint 11 [Overall Significance]), which is where **evaluation-specific logic and methodology** come into play. Later, in Chapter 10, we return to the KEC and show how all of the information we have covered fits into the big picture.

It is important to note that the KEC can be applied to a participatory or **collaborative evaluation** just as easily as it fits into the conduct of an independent evaluation being done for accountability. Whether the evaluation is a facilitated collaborative effort or not, the evaluation team members (be they external or **internal evaluators)** still need some guidelines for figuring out what should go into an evaluation to make sure that it provides the most accurate answers to the most important questions.

IDENTIFYING THE EVALUAND, ITS BACKGROUND, AND ITS CONTEXT

Before we plunge into the nuts and bolts of evaluation design, it is a good idea to first clarify what it is you plan to evaluate (i.e., your evaluand). This might seem like an incredibly basic question, but it trips up a lot of people. For your first evaluation, it is important to choose something manageable to which you could reasonably expect to gain access.

A clear and accurate description of your evaluand should appear under Checkpoint 2 (Descriptions and Definitions) of the KEC and should also have a brief mention in your evaluation report's Executive Summary (Checkpoint I). Equally important is to gain a solid understanding of the evaluand's Background and Context (Checkpoint 1). These three checkpoints are the focus of this chapter (Exhibit 1.3).

Exhibit 1.3 The Checkpoints Where the Evaluand, Its Background, and Its Context Are Described

I. Executive Summary
One- to two-page overview of the evaluand and findings

1. Background and Context
Why did this program or product come into existence in the first place?

2. Descriptions and Definitions
Describe the evaluand in enough detail so that virtually anyone can understand what it is and what it does.

When completing the Descriptions and Definitions checkpoint, the evaluation team should not just use brochures or Web sites to find out what the evaluand is *supposed* to be like; instead, the team should describe it as it *really* is. This usually involves, at a minimum, a firsthand visit and some interviewing of key stakeholders. The information presented under this checkpoint should be purely descriptive in nature; that is, you should not make comments here about the merits of the evaluand or its design.

At the same time, the evaluation team should conduct a preliminary investigation to find out what it was that led to the development of the evaluand in the first place and any underlying rationale for how or why it was intended to address the original need, problem, or issue. This information will go under the Background and Context checkpoint.

* TWO *

DEFINING THE PURPOSE OF THE EVALUATION

Having identified and described the evaluand, the first task in designing an evaluation is to determine its main purpose and the "big picture" questions that need to be answered. This feeds directly into Checkpoint II of the Key Evaluation Checklist (KEC), the Preface (Exhibit 2.1).

Exhibit 2.1 The Preface Checkpoint of the KEC

II. **Preface**
Who asked for this evaluation and why? What are the main evaluation questions? Who are the main audiences?

In the initial contact with the client, you will need to document who asked for this evaluation and why. As you scope the project, you should also be looking to clearly define the evaluation's purpose, its main audiences, and the big picture questions that need to be answered:

A. What is (are) the main purpose(s) of the evaluation?
   i. To determine the *overall* quality or value of something
   ii. To find areas for improvement iii. Both of the above

B. What is (are) the big picture question(s) for which we need answers?
   i. Absolute merit or worth (e.g., How effective/valuable/ meritorious is/was this? Is/Was it worth the resources [e.g., time, money] put into it?)
   ii. Relative merit or worth (e.g., How does it compare with the other options or candidates?)

Having a good understanding of both the broad purpose ("A" in the preceding outline) and the big picture questions that you need to answer ("B" in the outline) is a crucial first step before jumping into any kind of design. The evaluation methods you will use all hinge on getting this right. In the next few sections, we run through a few examples of when each purpose and type of question might fit the situation to help you get a good feel for this.

At the end of this chapter, you will identify the big picture questions that you need to answer about the evaluand and the primary audience for the answers. Note that the primary audience usually consists of more people than the person who initially asked for the evaluation. The various stakeholders you talk to may have some specific questions for you to answer, but they might not always be clear about which options under A and B in the outline apply. Your job as evaluator is to probe what it is that the organization really needs to know and to communicate this in a way that makes sense to your audience. (This usually means not using jargon such as *absolute* and *relative merit.)*

QUESTIONS ABOUT ABSOLUTE VERSUS RELATIVE QUALITY OR VALUE

Broadly speaking, there are two types of evaluation questions: those that ask about the quality (i.e., merit) or value (i.e., worth) of something in *absolute*[1]*;* terms and those that ask about the quality or value of something in *relative,* terms. Questions about the absolute quality or value of something (sometimes called **grading**[2]**)** may include the following:

- Is this intervention, product, or individual good enough to implement, buy, or hire (i.e., up to minimum requirements)?
- How should this level of performance be characterized (on a particular **dimension of merit** or overall)? Is it excellent? Good? Satisfactory? Poor? Completely unacceptable?

In contrast, questions about the relative quality or value of something are always asked in comparison with one or more other evaluands (e.g., interventions, products, job candidates). This evaluative activity is sometimes called **ranking,** even though it does not necessarily result in a strict numerical rank for each evaluand. Examples include the following:

- Which of these three pilot interventions is the most cost-effective and should be implemented throughout our organization?
- Who are the top 10% of our employees or students?

Now, how does this absolute-relative distinction link with the two main evaluation purposes (formative and summative) outlined earlier? It is easy to see how the preceding categories fit with summative evaluation. But even when the main purpose of the evaluation is formative, the absolute-relative

merit questions will still apply at least to *aspects* of the evaluand (e.g., performance on a specific outcome), if not always to the entire entity itself. Table 2.2 shows example evaluation questions under each category of the resulting 2x2 matrix.

Table 2.2 Example Evaluation Questions for Looking at Absolute Versus Relative Merit or Worth for Each Evaluation Purpose

Purpose of Evaluation	*"Grading " Questions*	*"Ranking" Questions*
Demonstrating or assessing overall quality or value (summative)	Is this national health intervention worth what it costs (in terms of time, money, and other resources)?	Which of these three executive development interventions we tried was the most cost-effective?
Finding areas for improving an existing evaluand (formative)	How well does the content of this training program (coverage, breadth, and depth) match the real needs of our minority trainees?	How do the initial improvements in manufacturing efficiency compare with those achieved elsewhere in the industry?

* THREE *

IDENTIFYING EVALUATIVE CRITERIA

In the first two chapters, we identified our evaluand, what "big picture" questions need to be answered about it, and who needs to know. Now it is time to roll up our sleeves and get into some of the nuts and bolts. One of the most important activities in putting together a solid evaluation is identifying the evaluative criteria or **dimensions of merit.** These are the attributes (e.g., features, impacts) of the evaluand that we will look at to see how good (or how valuable, how effective, etc.) it is.

The evaluative criteria are most relevant in five of the Key Evaluation Checklist (KEC) checkpoints: Consumers, where we identify who might be affected by the evaluand; Values, where we explain broadly how we define what is "good" (or what is "valuable"); Process Evaluation, where we evaluate the content and implementation of an evaluand; **Outcome Evaluation;** and Comparative Cost-Effectiveness. These checkpoints are reproduced in Exhibit 3.1.

Before we start exploring the strategies available for identifying evaluative criteria, it is worth spending a few minutes on the following question: Why not just use goals? After all, this is one of the most common strategies used by both managers and evaluators, that is, seeing whether the evaluand did what it was supposed to do.

WHY NOT JUST USE GOALS?

One of the first places many people start when they are asked to evaluate something is to find out what it was supposed to do and then check to see whether it did that. It is quite legitimate for management to want some information about performance relative to preset targets, and an evaluator is certainly the kind of person who has the expertise to collect such information. But as evaluators, we also must consider whether this information alone will allow us to draw valid conclusions about how well the product, project, or

Exhibit 3.1 The KEC Checkpoints That Are Most Relevant to the Identification of Evaluative Criteria

3. Consumers
Who are the actual or potential recipients or impactees of the program (e.g., demographics)?

5. Values
On what basis will you determine whether the evaluand is of high quality, is valuable, and so forth? Where will you get the criteria, and how will you determine "how good is good"?

7. Outcome Evaluation
How good or valuable are the impacts (intended and unintended) on immediate recipients and other impactees?

8 & 9. Comparative Cost-Effectiveness
How costly is this evaluand to consumers, funders, staff, and so forth, compared with alternative uses of the available resources that might feasibly have achieved outcomes of similar or greater value? Are the costs excessive, quite high, just acceptable, or very reasonable?

program is doing.

Before we get into a discussion of this, a quick point of clarification is in order. Most evaluands have some overarching purpose that we might refer to as a "goal." But that is not the kind of goal we discuss in this section. Rather, the term is used here to refer to the *specific objectives* that many evaluands have in place, complete with *preset targets* that might or might not be achieved.

Well-thought-out goals (in the sense of specific measurable targets to be achieved) can often take us part of the way toward working out how good (or how valuable, how effective, etc.) an evaluand is. Unfortunately, even the best ones have the potential to fall short in several important respects. Let's use an example to see why. Suppose that we had a hypothetical evaluand (called Program X) with three specific measurable goals. Suppose that Program X achieves one goal exactly, makes a near miss on another, but far exceeds target performance on its third goal (Exhibit 3.2).

Exhibit 3.2 Performance of Program X Against Its Three Specific Goals

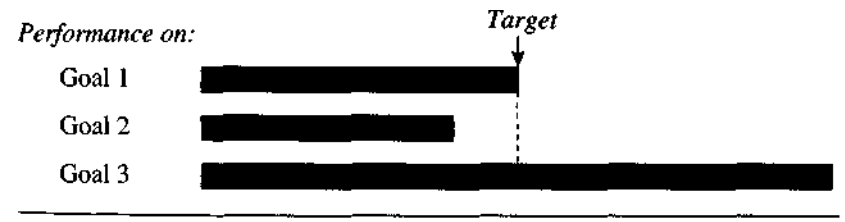


Table 3.1 lists some of the challenges encountered if one takes a strictly goal achievement-oriented approach to evaluating Program X.
The long and the short of it is that goals with specific targets can be handy guides when they exist but that even the best ones still need to be tweaked and/or supplemented with other criteria if they are to be used in an evaluation. What we really need is something more bulletproof that will allow us to take into account all of the issues listed previously and will even deal with the situation when there are no preset goals or targets or when we decide to do the evaluation in *goal-free* mode.

In **goal-free evaluation** (GFE), the evaluation team deliberately avoids learning what the goals are (or were) so as to avoid being overly focused on intended outcomes. The rationale behind this approach is that both intended and unintended effects are important to include in an evaluation. Therefore, it is important to find all effects, and it is of little consequence whether any identified effects happened to be intended or unintended.

Because the human mind inevitably pays more attention to what it knows it is looking for, concentrating on intended effects can lead the evaluation team to miss seeing some of the unintended effects. By leaving the search for effects (i.e., outcomes) open-ended and not focused primarily on goals (i.e., intended effects), GFE often picks up more side effects than does goal-based evaluation (GBE).

GFE is sometimes called **needs-based evaluation** because a **needs** assessment is one of the primary tools used to identify what effects (both positive and negative) should be investigated.

Table 3.1 Problems With Using Preset Targets or Goals as the Only Criteria

Problem	*Example*
Overruns and shortfalls	Should we (a) call Program X a "failure" because it missed one of its targets, (b) say that it did very well because it exceeded one of its targets by much j more than it missed on another, or (c) something in j between?
Goal difficulty	What if the goal that Program X barely missed (Goal 2) was a particularly challenging one, whereas the goal it far exceeded (Goal 3) was easy? |
Goal importance	What if the easy target was actually the more important one (i.e., it was more valuable to meet that goal than to meet the other one)? (How would we find out independently whether it was or was not? More on that in Chapter 7.)
Side effects	What if Program X also had an excellent side effect that was not included in the list of goals? Should we disregard that? If not, how would we know whether it compensated for the target it missed?
Synthesizing mixed results	What if we need to rank (or choose between) two programs, one of which is Program X and the other of which exactly met each of the three targets without exceeding any? On what basis could we say that one is better than the other?
Reasonableness of target levels	Suppose you find out that Program X came in "on budget" (i.e., met its cost goal), but then you find out that it cost five times as much as any comparable project that achieved roughly the same thing. (A similar problem occurs if Program X goes just over a very lean budget.)
Ignoring process: Do the ends justify the means?	What if Program X came in on budget by forcing the project staff to work overtime every weekend for 3 months so that in the end the top three team members quit their jobs and went to work for a competitor?
Whose/Which goals to use?	Suppose different people (program designers, management, and the staff who implemented the program) have very different versions of what the goals really are (i.e., what they are *really* trying to achieve). Whose/which goals should you evaluate against? *Related issues:* How will you handle the politics of choosing one set of goals over another? What will you do if you chew through your entire evaluation budget just finding out what the goals are?

NOTE: These points draw on Scriven's (1991) list of problems with goal achievement evaluation, with some adaptations and further explanations.

IDENTIFYING CRITERIA: BASIC CONCEPTS AND TOOLS

Identifying the right criteria for an evaluation is similar to deciding what symptoms to look at when determining what (if anything) is wrong with a Patient and how serious it is:

- We have a relatively limited time frame in which to make the diagnosis.
- If we miss something, we could easily make the wrong diagnosis.
- If we place importance on things that are not relevant to overall health, we could make an inaccurate diagnosis.
- Some *types* of symptoms represent more serious problems than do others.
- The *severity* of symptoms is important. Slight deviations from healthy levels are not as serious as those that are way off the mark.
- Sometimes it is *combinations* of symptoms that indicate a far more (or less) serious condition than each individual symptom would suggest.
- Sick patients are sometimes in denial about their symptoms or simply do not notice them, so there is a need to verify what they tell us (when possible) and to look for what they do not tell us.
- There could be several things wrong with the patient.
- In the end, we must put a lot of complex information together and come up with a final diagnosis (so that we know whether to admit the person to a hospital immediately or to send him or her home with some medication).

Unlike medicine, evaluation is not a discipline that has been developed by practicing professionals over thousands of years, so we are not yet at the stage

where we have huge encyclopedias that will walk us through any evaluation step-by-step. Even if we did, such "book knowledge" would not be enough. Like medicine, evaluation is an art and a craft as well as a science. Becoming a good evaluator involves developing the pattern-spotting skills of a methodical and insightful detective, the critical thinking instincts of a top-notch political reporter, and the bedside manner and holistic perspective of an excellent doctor, among many other skills.

For the beginner, this lack of structured guidance can be a real headache. Although doing evaluation in the real world involves many complexities, there are (thankfully) a number of fairly straightforward nuts-and-bolts tools that evaluators can use to get started.

When it comes to building a criterion list, there are a few tools and procedures that are either essential or very useful:

- A needs assessment
- A simple **logic model** that links the evaluand to the **needs**

- An assessment of other relevant values
- Checklists for thinking of other relevant criteria under the headings of Process, Outcomes, and Cost
- A strategy for organizing your criterion checklist

In the rest of this chapter, we run through what these tools and procedures are and how to use them. By the end, you should be able to draw up a good initial criterion list for whatever it is you plan to evaluate. Of course, you will often find that you need to tweak the list once you get into the evaluation proper because, for example, there may be some effects or issues that you did not anticipate. But the main thing is to go into the evaluation with a well-thought-out plan so that you know what you need to know, where to get that information, and how you are going to put it together when you write up (or present) your report.

For those readers with a particular interest in **policy evaluation,** the identification of criteria follows many of the fundamental principles described here but can be more complex in several ways. For example, one often must weigh very difficult conflicting values such as whether being able to freely choose a school for one's child is intrinsically valuable even if that choice leads to poorer educational outcomes for the child (Miron & Nelson, 1992). Colleagues who work in the areas spanning evaluation and public policy have recommended two books that complement the methods described here and help to span the gap between traditional **policy analysis** and evaluation. These are listed as Additional Readings at the end of this chapter.

NEEDS ASSESSMENT FUNDAMENTALS

The very basic idea behind needs assessment is as follows. Having a positive impact on end users (also referred to as **consumers** or **impactees)** is (or should be) the fundamental purpose that justifies the creation or existence of all products, services, programs, and policies in the first place. The primary *consumer* is the person or entity who buys or uses a product or service, enrolls in or is the recipient of a program, is directly affected by a government policy, and so forth. (There may also be some others affected indirectly or unintentionally, hence the use of the more inclusive term *impactees.)*

If we can understand what the true needs of consumers or impactees are, this gives us a solid basis for finding out how well a program is doing by seeing how well it is helping to meet those needs. In other words, needs that we identify become the outcome criteria we use for the evaluation. Furthermore, ′ the data collected during the needs assessment phase can often double as **baseline data** if we wish to track change in certain outcome variables.

Before launching into a needs assessment, then, a useful first step is to figure out who our consumers or impactees are. We had a shot at this in the exercise at the end of Chapter 1, but let's clarify a few key points and double-check to make sure that we have this right.

IDENTIFYING CONSUMERS OR IMPACTEES

> In general, consumers (or impactees) are those people for whom something changes (or should or might change) as a result of a particular product, service, program, or policy.
>
> Occasionally, products, services, programs, or policies are designed to prevent change rather than to effect change (e.g., cosmetics that slow or prevent the signs of aging). In such cases, the impacts, effects, or outcomes are the lack of change that otherwise would have occurred and the impactees are the users or receivers of the products, services, or programs.

Recall the KEC, where Checkpoint 3 states that an important part of evaluation is to correctly identify the consumers. When we talk about con- : sumers or impactees, we are referring to those people for whom something changes (or should or might change), or for whom something is prevented from changing, as a result of our product, program, or policy. At this point, we do not include the **upstream stakeholders** (i.e., the people who worked on the design, implementation, and/or management) under the heading of consumers or impactees; rather, we include just the people "downstream," that is, **down-**

stream consumers (Exhibit 3.3). **(PAGE 31)**

Consumers can be divided into two groups: (a) immediate users or recipients and (b) other downstream impactees (Exhibit 3.3). **Immediate recipients** are the people who actually bought a product, signed up for a program, or received services directly from the evaluand, whereas downstream impactees are those who were not **direct recipients** but who were affected nevertheless. Downstream impactees need not be individuals; instead, they could be the unit or organization where direct recipients work, the local community, or society in general.

Table 3.2 gives some examples of the different types of consumers for programs, policies, and products. When listing consumers, you should always include not only those who actually received the product, service, or program but also those who potentially could have or should have done so. This is because the extent to which a program or service actually reached those who most needed it is part of what makes a good program or service. If we consider just the impact on those who happened to be reached, we might be missing a big chunk of the story. In product evaluation terms, it may be helpful to think of the categories of immediate consumers as the "potential target markets."

Table 3.2 Examples of Consumers Identified for Different Evaluands

Evaluand	*Immediate Recipients (actual or potential)*	*Downstream Impactees (actual or potential)*
After-school chess program	Children who attended the program; other children in the area who do not currently attend other after-school activities	Siblings and families of children who attended the program; the local community
Executive coaching intervention	Executives who received coaching; other executives or managers within the company who did not receive coaching	Executives' direct reports; the senior management team; the chief executive officer; shareholders; the organization as a whole
Policy to decrease legal drinking age from 21 to 18 years	People 18 to 20 years of age	Parents and siblings of 18- to 20-year-olds; the police; bar and restaurant managers; the general public (especially those who patronize bars)
New 1-kg (2-pound) lightweight portable printer	Professionals and academicians who are frequent travelers or who have a "just-in-time" approach to doing presentations; people in small apartments or dormitories	Colleagues; clients; the organizations where primary consumers (immediate recipients) work
Farm irrigation project	Farmers who received irrigation; farm workers; farmers and other landowners in the area who did not receive irrigation	Adjacent landowners; produce vendors; the surrounding community (e.g., people who consume farm produce, local businesses)

Needs Versus Wants

There are two critically important things you must know to design a good needs assessment. One is the fundamental difference between wants and needs. The other is what the distinctions are among the different kinds of needs, not all of which we are concerned with in a needs assessment.

Importantly, a true need might *not* be something that someone desires or is conscious of needing. It might even be something that is definitely not wanted. A seriously dehydrated person wandering in the desert might strongly desire a beer on arrival at an oasis, but what he or she really needs is water.

A *need* is something without which unsatisfactory functioning occurs.'

In contrast, a *want* is a conscious desire without which dissatisfaction (but not necessarily unsatisfactory functioning) occurs.

Let's try a more complex example to demonstrate the distinction between wants and needs. If you ask 14-year-olds whether they really need to know how to do algebra, they might tell you that it is one of life's crudest inventions and a completely unnecessary one at that. What they are expressing in this case are wants and not needs. It is a fact that, in virtually all societies around the world, some understanding of algebra (and mental arithmetic) is essential to make sure that one does not get "ripped off' when buying timber for building a house or when buying food for one's family. Getting ripped off is clearly unsatisfactory functioning, so a certain level of knowledge in algebra and arithmetic is a need.

The Context Dependence of Needs

Another important point here is that needs are highly context dependent, with context having many dimensions such as geographical, cultural, and historical. A century ago, we did not need to be able to get from one side of the Pacific Ocean to the other in less than a day, but in today's business environment, that is expected. A firm doing business overseas would fall way behind its competitors (a clear example of unsatisfactory functioning) if representatives traveled only by sea and took weeks to get to their overseas customers. The context has changed, and the need has changed with it. As another example, basic living condition needs are defined differently in different countries because "satisfactory functioning" is defined differently.

If needs are context dependent, does that mean that they all are arbitrary? Not at all. Common sense and good evaluation practice dictate that we need to clearly define the context and justify why we classify certain things as needs. If there is disagreement on this, so much the better. It can help to spark an important dialogue about how "need" should be defined in a particular context, and this is an extremely important conversation to have.

This is not meant to imply that you, as an evaluator, are somehow infallible. Evaluation is a tough job that is difficult to get right, and you should *always* be open to the possibility that you have missed something, incorrectly assumed something, included something irrelevant, or made some other blunder. What we try to do throughout this book is make the evaluation's methods and findings as **systematic** (step-by-step and thorough), **objective** (free from unacceptable bias), and **transparent** (easy to follow) as possible. This makes it easier for you and others to pinpoint exactly where you might have gone wrong as you drew your evaluative conclusions.

Different Kinds of Needs

We have distinguished needs from wants. Now we need to make sure that we understand the different kinds of needs. Basically, there are three dimensions on which we can distinguish needs.

The main dimensions that distinguish the different kinds of needs are as follows:

1. Conscious needs versus unconscious needs
2. Met needs versus unmet needs
3. Performance needs versus instrumental needs[2]

The distinction between *conscious needs* and *unconscious needs* is a fairly straightforward one—the things we know we need versus the things we do not know we need. And, as pointed out earlier, there are things that we *think* we do not need but that we actually do need. The term *unconscious* is not meant to imply that these needs are not known to anyone; rather, it just implies that the needs are not known to the person who has the needs.

A trickier distinction is that between *met needs* and *unmet needs.* The idea here is that just because someone already has something does not mean that he or she does not need it. Suppose that a group of rural farmers has good irrigation to their crops. Does this mean that irrigation is not needed? It is true that irrigation is not an unmet need in this case, but it is certainly something that, if taken away, would probably cause seriously unsatisfactory functioning, including possible crop loss.

Why bother with looking at needs that are already met? Whatever we are evaluating is designed to address unmet needs, right? Yes, but do not forget unintended consequences. A good evaluand is something that not only adds good things (e.g., services, products, opportunities) but also does not take away something important in the process. For example, building a new factory in an economically depressed town may provide employment for local people, thereby addressing an unmet need. But what if it also drains most of the town's water supply and/or seriously pollutes the air, thereby taking away the previously met needs of clean air and water? Evaluation involves not only looking at how well problems (unmet needs) were addressed but also looking at whether any new problems or benefits were caused.

The third (and most difficult) distinction is that between ***performance needs*** and *instrumental needs.* A performance need is a state of existence or level of performance that is required for satisfactory functioning. Roughly, it is a "need to do" something, a "need to be" something, or a "need to be able to do" something. In contrast, an instrumental need is the product, tool, or intervention that is required to address the performance need.

If we say that traveling executives need lightweight laptop computers, that is an example of an instrumental need. If we say that these executives need to be able to access e-mail and files while on the road, that is the performance need. The important thing to notice here is that the performance need is a lot easier to argue as a defensible fact than is the instrumental need. After all, one could also access e-mail and files through a handheld computer or personal digital assistant (PDA) or by using business centers or Internet cafes. If an executive has possession of or access to one of these, he or she might not need a laptop at all.

In short, the performance need is the *actual or potential problem,* whereas the instrumental need is the *proposed solution.* In needs assessment, we are concerned with the performance needs and not the instrumental needs.

As we will see, this has major implications for needs assessment methods.

NEEDS ASSESSMENT METHODS: A TWO-PHASE APPROACH

Needs assessment, as conceptualized here, consists of two phases:

1. Identifying and documenting performance needs (severity documentation phase)
2. Investigating the underlying causes of performance needs (diagnostic phase)

IDENTIFYING OTHER RELEVANT CRITERIA

Once you have identified the main needs for the program, your next step is to think through what other considerations might be relevant to this evaluation. Table 3.4 lists the main possibilities that should be considered in addition to needs (Scriven, 2003) and shows how they would be applied to a grantsmanship workshop with follow-up technical assistance.

The list of criteria in Table 3.4 should be kept alongside the list we generated from the earlier logic model-based needs assessment. You may have noticed some overlap between the two. That is not a problem. Together, the two lists will form the main ingredients for generating a complete list of criteria under the headings of Process Evaluation, Outcome Evaluation, Comparative Cost-Effectiveness, and Exportability (KEC Checkpoints 6-10). We address this in the next chapter. **[TABLES AT PAGES 48-49]**

* FIVE *

DEALING WITH THE CAUSATION ISSUE

No discussion of evaluation nuts and bolts is complete without some mention of the **causation** issue. Although this is a relatively simple concept to grasp in everyday life, causation is both one of the most difficult and one of the most important issues in evaluation. Even if we observe changes that are consistent with the expectations or goals of a program or another evaluand, we cannot correctly refer to these as "impacts" or "outcomes" unless we can demonstrate that the evaluand was at least a primary cause of those changes.

Strategies for inferring causation form a key part of what should be written into the Methodology checkpoint of the Key Evaluation Checklist (KEC) (Exhibit 5.1). The choice of evaluation design affects the evaluation team's ability to make causal inferences. Causation is also relevant for the Outcomes checkpoint because identifying anything as an outcome is saying that it was caused by the evaluand.

One of the great challenges with causation is that the further down the causal chain (toward what we might call "ultimate outcomes") one goes, the more other factors come into play. For example, the career success of a university graduate can be attributed not only to the quality of education he or she received but also to the quality of mentoring and advancement opportunities after graduation, support from family, aptitude or intelligence, and many other factors. The fact that so many variables are in play makes it quite difficult to pin down whether career success (and other downstream changes) is substantially due to the evaluand (in this case, the program from which the student gradu¬ated) or can be attributed mostly to other factors.

Although the causation issue is incredibly important, demonstrating causal links can seem like an impossible task, especially for evaluators with limited time and resources (most of us likely fall into that category). For this reason, many people abandon the issue altogether, either by tacking on a bunch of disclaimers to their evaluations or by downplaying the importance of causal analysis.

Here is the good news: There is some practical light at the end of the cau¬sation tunnel, and the tunnel is not nearly as long and treacherous as legend has it. To deal with the causation issue, we need to answer four

important questions in the following order:

- How certain does the client need us to be to say that the evaluand "caused" a certain change?
- What are the basic principles for inferring causation?
- What types of evidence do we have available to help us identify or rule out possible causal links?
- How should we decide what blend of evidence will generate the level of certainty needed most cost-effectively?

Exhibit 5.1 The KEC Checkpoints for Which the Causation Issue Is Most Relevant

III. Methodology
What is the overall design of the evaluation (e.g., quasi-experimental, participatory, goal free)? Explain why (briefly).

7. Outcome Evaluation
How good or valuable are the impacts (both intended and unintended) on immediate recipients and other impactees?

CERTAINTY ABOUT CAUSATION

Many readers of this book may have noticed that in the academic literature, research conclusions are often so laced with disclaimers about causation that one wonders whether it is possible to demonstrate causal links at all. Why would an evaluator on a limited budget even bother trying? The trick here is to understand two things. First, there are some major differences between the standards of proof being used by academics and what may be appropriate for us to use. Second, many of the methods used to address causation in the empirical literature are, quite frankly, pretty weak on the causal inference front.

Every profession has its own "dialect," including special terminology and rules regarding how to talk about things. For academics in the hard sciences and at least most of the social sciences, the norms dictate that even if researchers have evidence that makes them 99% sure that something is true, they still cannot say that they "know" or have "proved" it. Instead, the language is always framed in a cautious way, for example, "The evidence appears to suggest. . ." or "We found tentative support for . . ." This is in sharp contrast to the way in which we (and our clients) use terms such as *know* and *certain* in everyday conversation.

Organizational reality in for-profit, not-for-profit, and many government settings is that most decision makers would say that they *knew* something—and were prepared to make decisions on the basis of that knowledge—if they were, say, 70% or 80% certain based on the evidence. Of course, this varies a bit from setting to setting and from decision to decision, but most would agree that this sounds about right.

Our task as evaluators is to provide timely answers about the quality or value of products, programs, policies, and other evaluands, often to help people make decisions. These may be internal decisions about how to improve something or consumer decisions about which product to buy or which school to attend. Because each decision-making context requires a different level of certainty, it is important to be clear up front about the level of certainty required. Then, rather than throwing in the methodological kitchen sink or skipping the causal inference step, we will be in a much better position to strategically put together a blend of methods that will meet that certainty requirement (Davidson, 2003).

Some of the research in the academic literature tends to be somewhat lacking in evidence for making causal inferences. There are two reasons for this. One is that many researchers use wholly quantitative or wholly qualitative methods in their studies. The other reason is that in quantitative studies, researchers often lack the opportunity to use large samples, control groups, and random assignment. In such cases, quantitative methods alone tend to be woefully inadequate for attributing causation, as are many all-qualitative designs. So, those disclaimers about causation that we see in such single-method (i.e., all-quantitative or all-qualitative) research are almost certainly justified. Moreover, they are attributable to shortcomings in the research design itself rather than to the impossibility of solving the problem.

INFERRING CAUSATION: BASIC PRINCIPLES

What are we trying to do when we infer causation? There are two basic principles here. First, look for evidence for and against the suspected cause (i.e., the evaluand). Second, look for evidence for and against any important alternative causes (i.e., rival explanations).

When considering whether the evaluand caused the observed changes, the evaluation team members need to consider what evidence, if present, would help to convince them that this was the case. Conversely, if such evidence were absent, to what extent would that convince them that the evaluand was probably not the cause?

Equally important in causal analysis is the careful consideration of any and all important rival explanations for the observed changes. But how do we know which rival explanations are most important and how many we need to eliminate? That all depends on what level of certainty you need in your decisionmaking context. Sometimes you will need to eliminate only the "primary suspects," that is, the most likely alternative explanations. Sometimes you will need to rule out just about anything that anyone can

suggest.

Probably the best way in which to approach this task is with a stepwise process. The first step is to put yourself in the shoes of the harshest critics you can imagine and think what objections they might raise to your claim that the evaluand caused a particular effect. Using a mix of strategies from the next section, gather enough evidence to confirm or rule out that rival explanation. Then consider what the next objection is likely to be. Repeat the process until all remaining alternative explanations are unlikely enough that they do not threaten your conclusions given the level of certainty needed to make them.

Usually, the more politically charged or controversial something is, the more likely it is that there will be opponents, many of whom will attack the methodology of the evaluation if they do not like the conclusions. And the harder people attack, the more solid your answers need to be. For this reason, the level of certainty required may change depending on what you uncover in the evaluation.

Even if a fairly high level of certainty is required, the trick is not to focus on a single "Rolls Royce" method for causal inference (e.g., an elaborate experimental design with multiple controls). Rather, you should use a strategic mix of methods that have different strengths and that *together* will give you enough evidence to be certain enough that the link is (or is not) causal. This principle (using methods with different strengths to complement each other) is called *critical multiplism* (Shadish, 1994).

INFERRING CAUSATION: EIGHT STRATEGIES

Some academics, among others, frequently say that the only way in which to infer causation is with the use of randomized **experimental designs.** This is sometimes met with a response from practitioners that such methods are simply not feasible in real-world settings. This is not true.

There is good news on both fronts for evaluators who need to know whether the changes they are seeing really are outcomes (i.e., changes attributable to the evaluand)—and that is all of us. The fact of the matter is that experimental designs (or at least **quasi-experimental designs)** are actually quite viable more often than we might expect. But even when they are not, there are several other practical strategies, some of which make use of some very powerful qualitative methodologies, that can be used to supplement or even replace the use of experimental and quasi-experimental designs.

The following subsections describe a range of methods for inferring causation, from very simple commonsense strategies to some more complex methods. For a small-scale evaluation, even some modest evidence about causation could prove to be sufficient. For more high-stakes evaluations, the evaluation team will need to draw on a range of methods to attain the level of certainty required.

Strategy 1: Ask Observers

Suppose that someone asked you to name the four or five most important factors that led to the development of your current professional skill set. Most of us could easily identify which experiences were the most important and which ones had nearly no effect whatsoever. For the powerful learning experiences we have in our careers, there is no doubt in our minds that the experiences were primary *causes* of the learning. In many cases, we can also identify important contextual factors or we can point to a combination of experiences that culminated in a quantum leap in knowledge. And we can just as easily list a number of courses, books, conferences, and work assignments that added very little (i.e., where there was virtually no causal link).

It is amazing how "arm's length" we are as we look at the impacts of things on people's lives, especially in quantitative research. We gather pre- and posttest measures, and then use regression and other statistical tools to partial out the extraneous effects of this and that, without ever considering that perhaps we should start by just asking the question directly. In qualitative research, such evidence is perhaps more likely to be collected, but it is often not treated as explicit evidence of causation.

The "ask observers" strategy includes two possibilities. The first is to directly ask people who were supposedly affected by the evaluand (i.e., actual or potential impactees). The second possibility is to ask those who were in a position to observe the effects on impactees (e.g., coworkers, parents, teachers, trainers).

There are two ways in which to infer causation by just asking the people who were supposedly affected by the evaluand. One is to first gather some data about changes in outcome variables (e.g., reduced absenteeism, improved performance) and then to identify those people who experienced (a) little change, (b) some change, and (c) substantial change (positive or negative). In a follow-up interview or survey, the evaluation team members could ask, for example, "We noticed that you have had a substantial decrease in the number of times you were absent from or late to work during the past few months. Can you tell us a little about why that is?" The answer would tell you whether the individual believed that the evaluand was the primary cause or not and/or the extent to which other factors (e.g., contextual factors, other events) might also have contributed to the change. Note that the use of an open-ended question here allows respondents to list other causes that the evaluation team might not even have considered.

The other way in which to gather causal information from those directly affected by the evaluand is to actually work causation into the survey or interview questions themselves. So, instead of asking people to rate their level of knowledge before and after completing a training or educational program, you might ask directly, "How much has your knowledge increased *as a result* o/participating in this program?" (Include the italics in the survey item to make sure

that respondents pay attention to it.) To probe other causes of knowledge gain, you might ask, "Did anything else besides the program increase your knowledge in this area over the same period of time?" To get at side effects, you might ask, "Please describe anything else that has happened to you or someone you know *as a result of* participating in this program." This way, you are not simply asking what has changed since before the program; instead, you are asking directly about the tilings that people know or believe were caused by the program.

Some researchers may argue that causation-rich questions such as these are leading, that is, that they implicitly direct the respondent to answer in a particular way (usually positive). It is true that we need to be careful about question wording when designing interview instruments or questionnaires, bearing in mind that in most cases, it is quite obvious what the evaluation team is trying to get at. But do not forget that these same questions, if well constructed, can also provide the opportunity for the respondent to say, "My knowledge of X increased during that time, but not because of that program." The arm's-length pre- and postquestionnaire that does not ask about causation eliminates the opportunity for people to even mention this.

A great example that incorporates both of these "just ask people" strategies just described is Brinkerhoff's (2003) Success Case Method. All participants in a particular program are given a 5-minute questionnaire on which they are asked whether or not they have been able to achieve enhanced performance as a result of the program and, if so, to give an example. Claims of dramatic improvement are then cross-checked against hard data to identify the true success cases, and a sample of these individuals are then interviewed in-depth to find out what it was that allowed them to get so much out of the program. In this case, the causation question is not just *whether* the program produced the effect but also *what other factors* enabled or inhibited the effect.

Some might argue that the individual might not be a reliable witness to help answer the causation question. In rare cases, this may be true. However, there are few evaluands that are so subtle in their effects that the recipient does not even notice their influence, so it seems remiss to exclude the views of the very people who likely saw things happen with their own eyes or experienced change directly. Of course, in most cases, other evidence will also be required to make justifiable causal inferences.

The "ask observers" method is not limited to those who were themselves changed. Often it is possible to identify people who observed a cause produce an effect in someone (or something) else. For example, parents of very young children can often directly observe the influence of particular experiences on their children (e.g., whether children mimic violent acts after watching a certain television show). A spouse might be in a good position to observe whether a violent offender's behavior was affected by a counseling session. Or an observer in a mathematics class might be able to see directly whether children learn faster and are more engaged when a new practical exercise is used to illustrate a concept.

Strategy 2: Check Whether the Content of the Evaluand Matches the Outcome

Here is another super simple commonsense strategy for inferring causation. Suppose that a treatment program for alcoholics taught participants several very specific strategies they could use to avert potential relapses. Also, suppose that participants in this program really did have very few relapses after completing the program. If the program were truly the cause of the lack of relapses, the evaluation team would expect to find that the alcoholics who avoided relapses used the strategies they had been taught in the treatment program rather than other strategies they knew previously or had picked up elsewhere. In other words, the content of the evaluand should quite often be reflected in some of the outcomes themselves if the evaluand did indeed cause the observed change.

When using this method, it is equally important to look for counterexamples. In this case, that means other strategies that were not learned in the program but that were used successfully to avert relapses. Where (or from whom) were these strategies learned? This information may point to one or more additional causes of alcoholics' success that were not attributable to the specific program. The existence of these additional causes does not negate the value of the program. However, if all potential relapses were prevented using strategies other than those taught in the program, and especially if relapses were not prevented in several cases where the taught strategies were used, this would call into question the value of the relapse avoidance strategies (and perhaps of the entire program).

Strategy 3: Look for Other Telltale Patterns That Suggest One Cause or Another

In addition to looking to see whether early outcomes (e.g., the behavior changes just described) match the content of the evaluand, it is often possible to identify other telltale patterns that suggest a particular cause. These patterns, or "signature traces," are described by Scriven as the key to making causal inferences using the *modus operandi method.* This method uses the detective metaphor to describe the way in which potential causal explanations are identified and tested. Scriven describes how chains of causal events often leave signature traces that the evaluator tracks down by moving both up and down the causal chain. Starting with the observed effects, or "clues," one can move up the causal chain, identifying what might have caused them.

In the opposite direction, one can start with the evaluand itself, or the "suspect," and trace down the causal chain to see what impacts it might have had and through what mechanisms. If evidence is consistent with the expected "trace" left by a particular causal chain, confidence in that chain as the correct

causal explanation is increased. Evidence that contradicts the expected trace eliminates that causal chain as a possibility, and missing evidence makes the explanation more doubtful.

The modus operandi method works best for evaluands that have highly distinctive patterns of effects. For example, a faith-based marriage counseling program, if effective, not only would result in partners using strategies taught within the sessions to improve their marriages but also would be likely to yield a telltale pattern of distinctive side effects. We might expect participants to report increased spiritual enlightenment and stronger connections with the relevant faith community. We might also expect to see less tolerance of attitudes and behaviors that are inconsistent with participants' faith. In contrast, improvements in marital relationships that were due not to the faith-based element but rather to the regular counseling would not be expected to yield such a pattern.

In some cases, there is not a great deal known about the patterns we should expect if a certain evaluand is likely to cause a particular effect. In such cases, it can be useful (albeit a weaker option) to draw an analogy with what is known about something similar. So, if the pattern observed closely resembles a known pattern in an analogous case, this can be interpreted as at least partial evidence for a causal link.

Let's use an example to illustrate the use of analogy as partial evidence of causation. Suppose that you had been asked to evaluate a cutting-edge intervention that helped teams of people to critically reflect on their work and generate new ways of doing things. Also, suppose that there was virtually no documentation about what happens when such interventions are successful. As an alternative, the evaluation team might dig for similar" interventions that had not been used on teams. Previous research shows that when individuals are taught to critically reflect on their own work (e.g., in executive coaching), they can make transformational improvements in their; own performance. The team learning intervention seeks to translate this' idea for an interactive team setting. By examining the'patterns in executive; coaching success cases and seeing whether they are mirrored in the team; intervention, it may be possible to use this as indirect evidence for a causal link by drawing an analogy with the individual-level version of the intervention. This evidence alone will not allow the evaluation team to make ; causal inferences, but it is certainly one additional piece of evidence to add to the pool.

Strategy 4: Check Whether the Timing of Outcomes Makes Sense

In nearly all cases, an outcome should appear only at the same time as or after whatever caused it.[1] With distal outcomes in particular (i.e., those quite far downstream in the causal chain), the evaluation team should expect a considerable delay between the introduction of the evaluand and the appearance of outcomes. In general, the further downstream the outcomes, the longer they should take to appear.

For example, suppose that we were evaluating a community health intervention that focused on improving diet and exercise. We should probably expect to see the following:

- Fairly immediate knowledge and skill gain relating to the subject matter taught as part of the intervention (i.e., we should be able to detect this during and/or immediately after any health education component)
- A short delay (days to weeks) before the knowledge and skills are translated into changed behavior such as improved eating habits and exercise
- A moderate delay (weeks to months) before we could expect to see changes in individual health indicators, such as cholesterol, weight, and blood pressure, as a result of sustained behavior change
- A long delay (probably years) before these changes could be expected to have become widespread enough in the community to affect community-level health statistics such as the incidence of diabetes and heart disease and average life expectancy

Information about expected time frames for outcomes may be found in the relevant literature and from experts in the field. But in many cases, the evaluation team's logic might not be too far off target, so just taking the time to think through the timing issue will probably pay dividends.

There are three ways in which this information can be used to help confirm or disconflrm causal links. First, each identified outcome should be checked to ensure that it did not occur either before the evaluand was introduced or unrealistically quickly afterward. In fact, this is one good reason to check on some of those downstream outcomes at points in time when it should be too early to detect any change.

Second, outcomes should also be checked to see whether the timing of their appearance would be more (or equally) logical relative to other possible causes. For example, suppose that on-the-job performance improved following a well-executed training program that also coincided with the introduction of a performance-linked bonus system. In this case, the evaluation team would look at the timing of the improvements relative to the introduction of the two interventions to try to work out whether one or both of these (and/or something else) were likely to have been a substantial cause of the improvement.

The third strategy for using information about the timing of outcomes is to check whether outcomes further downstream in the logic model did not occur out of sequence, that is, before the outcomes that were expected to lead to them. In the earlier example of a community health program, if participant

cholesterol and blood pressure dropped prior to any change in eating or exercising behavior, this makes it unlikely that the observed improvements in health indicators were caused by the program.

For those readers interested in exploring the timing of outcomes in more depth, Lipsey (1989) presents a very useful set of graphs that show different patterns of responses to interventions, including a delayed reaction and an initial response followed by a decay.

Strategy 5: Check Whether the "Dose" Is Related Logically to the "Response"

In the messy real world of evaluation, we are often faced with situations; where an evaluand has been implemented inconsistently. For example, th author was once asked to evaluate the effectiveness of a new **management-by-objectives** (MBO) and reward system. A year after the system was rolle out organization-wide, it turned out that approximately a quarter of all stj still had no objectives in place and that the range in the quality of performanc objectives was extremely variable across the organization for those who ha objectives in place. Although the social scientists in us*might throw our hand in the air in frustration in this kind of situation, the shrewd evaluators in u should instantly spot this as an excellent opportunity to check the causal link! between the evaluand and its suspected effects.

The dose-response idea (i.e., if more A, then more B) comes from" the medical metaphor of drug testing—the higher the dose, the greater the. response should be (up to a point). For a performance management system[5] such as the one just described, the more completely and effectively the system) had been implemented in a particular work unit, the higher the "dose" (of • MBO) for that unit and the greater the expected improvement in performance.' If we found that performance had improved more dramatically in units where the system had been poorly implemented (or not implemented at all), this; would be evidence that the new performance appraisal system was probably not the cause of the improvement.

When looking at the relationship between the "dose" of the evaluand and the "response" (magnitude of the outcome), it is important to bear in mind that this might not necessarily be a linear relationship. It is very common to have a "ceiling effect" where longer duration or more intensive exposure starts adding little or nothing in incremental value beyond a lower dose. Also, in many cases, there might be an "overdose" where excessive exposure or duration backfires and produces a less than optimal (or a very negative) result. As a simple example, schoolchildren will probably tolerate only so many hours per week of extracurricular reading before they develop a loathing for the activity.

An extension of the dose-response relationship is the situation where multiple doses are given and multiple responses are observed. Evidence for causation is strengthened if the evaluand is implemented in several different contexts and if the effect is observed every time (or nearly every time) the cause is introduced (i.e., when A, always B).

Strategy 6: Make Comparisons With a "Control" or "Comparison" Group

The dichotomous (on/off) version of the dose-response relationship is the comparison between people who have been recipients of an evaluand and those who have not. This relationship forms the basis for the classic experimental design. In a fully randomized experimental design, participants would be randomly assigned to either a treatment group (receive the evaluand) or a control group (receive nothing or an alternative intervention). Provided that sampling is done carefully and that sample sizes are large enough, randomization helps to make sure that there are no systematic differences between the evaluand recipients and nonrecipients. It is rather like thoroughly shuffling a deck of cards to minimize the chance that one player gets all of the high cards.

In a quasi-experimental design, groups would not be randomly assigned, but the evaluation team would seek out a closely similar comparison group with which to compare results. Careful matching of treatment and comparison groups eliminates or greatly reduces the likelihood that rival explanations exist (e.g., the groups were different from the start). For example, studies of the effectiveness of the death penalty have compared crime rates in adjacent counties across state lines where one state introduces or abolishes the death penalty but the other state does not. Researchers carefully check to ensure that prior crime rates are similar and that the inhabitants of each county are similar demographically, socioeconomically, and in any other important respects to make sure that the comparison is reasonable to make.

Strategy 7: Control Statistically for Extraneous Variables

In the statistical analysis of data from experimental, quasi-experimental, and even single group (dose-response) designs, it is often possible to "control for" certain characteristics of the recipients and/or the contexts that are suspected of being correlated with the outcomes. This is particularly useful in cases where the evaluation team cannot be certain that the control or comparison group (if any) is truly similar in these respects.

For example, suppose that you were evaluating an innovative new method for teaching mathematics in a high school and that you had decided to use a comparison group of classes that were not exposed to the new technique. Even if you were able to randomly assign students to the classes that used and did not use the method, it might still be useful to make sure that prior aptitude in math was not causing the results to look better or worse than they really were. The simple way in which to check this is to compare the

treatment and control [classes on prior math performance or scores on an aptitude or achievement test to ensure that there was no significant difference. But another more sophisticated strategy is to use a statistical technique called *regression analysis* to "partial out" the effect of prior aptitude in math so that any differences observed were not due to that factor. In this way, the evaluation team can statistically control for characteristics that might cloud the results.

Options and strategies available in the area of experimental, quasi-experimental, and related designs and associated data analysis are very numerous indeed. For some evaluations, these designs are essential. In such cases, if the evaluation team members do not have a specialist to help with the design, they j would be well advised to find one. And in the meantime, there are many , resources available to give the beginner a simple overview of the principles t and enough know-how to design simple experimental studies.

Strategy 8: Identify and Check the Underlying Causal Mechanism(s)

Another commonsense strategy we use a lot in everyday life is to look for an underlying mechanism that will help to make the case for causation more or less convincing. For example, the link between cigarette smoking and lung cancer was for years argued to be purely correlational. However, when research identified several substances known to be carcinogenic in cigarette smoke, it became more difficult to argue that there was not a causal link.

As a second example,[2] suppose that a team of consultants had been brought into an organization to facilitate a team learning intervention. The organization has shown an increase in profitability for the past quarter, and management wants to know whether this was due to the team learning intervention or to something else. How might the evaluation team use causal mechanisms to trace potential causal links?

The logic model in Exhibit 5.2 shows how this hypothetical team learning intervention would probably affect the bottom line. Evidence in favor of a causal link would include (a) an increase in investigation and critical dialogue skills during the intervention and (b) evidence that cost-saving improvements were identified or implemented during the intervention itself. Note that the logic model also includes the important contextual factor of a supportive work environment, which would be required for the success of the intervention.

Evidence against a causal link would include (a) no evidence of improved investigation or critical dialogue skills, (b) no evidence that the intervention had a motivating effect (with most participants complaining that it was boring), and (c) most employees attributing their improved performance to the new incentive system rather than the team learning intervention.

Where would a logic model like this come from, and what would make it more or less useful as a source of evidence for causal inference? An evaluation team with knowledge of team learning interventions (and access to the relevant literature) would be able to create a model that is consistent with cutting-edge knowledge about team learning.

* SEVEN >

DETERMINING IMPORTANCE

A frequent argument from those who *oppose* the notion that part of the evaluation team's job is to be explicit about quality, value, or importance is that no valid methodologies exist for doing so (Lawler, Seashore, & Mirvis, 1983). It is true that the average research methods text leaves the reader pretty well in the dark on this topic. But it is equally true that there has been significant headway made on the evaluation-specific methodologies available for the tasks of importance weighting, merit determination, and synthesis.[1] That is where we are headed in this chapter as well as the next few chapters.

Importance determination is defined here as the process of assigning labels to dimensions or components to indicate their importance.

When referring to importance determination, the term **importance weighting** is sometimes used. Conceptually, this is reasonably accurate. However, it does tend to make people think immediately of numerical weighting systems, which comprise only a small slice of the possibilities here. Whether one uses numbers, words, or symbols to signify importance matters little until we get to the synthesis step.

Importance determination is most relevant to the Sub-evaluation checkpoints and to the Overall Significance checkpoint of the Key Evaluation Checklist (KEC) (Exhibit 7.1). Under the Sub-evaluation checkpoints, the evaluation team ' needs to determine the relative importance of the various aspects of the evaluand investigated in addition to determining the merit of performance on each of those aspects (this is covered later, in Chapter 8). Under the Overall Significance checkpoint, all of these strengths and weaknesses are combined together based on their relative importance to draw overall conclusions. Methods for combining these are covered in Chapter 9.

DETERMINING IMPORTANCE: WHAT AND WHY

As we look down any list of evaluative criteria, it is intuitively obvious that not all of the criteria are equally important. The same is true when looking at the performance of an evaluand across various components. Knowing which criteria and/or components are more important is essential for being able to (a) prioritize improvements, (b) identify whether identified strengths or weaknesses are serious or minor, and/or (c) work out whether an evaluand with mixed results is doing fairly well, quite poorly, or somewhere in between. In this section, we examine the distinction between **dimensional evaluation** and **component evaluation** as well as how it affects the importance determination task.

Exhibit 7.1 The KEC Checkpoints Where Importance Determination Is Used

6. Process Evaluation How good, valuable, or efficient is the evaluand's content (design) and implementation (delivery)?	**7. Outcome Evaluation** How good or valuable are the impacts (both intended and unintended) on immediate recipients and other impactees?	**8&9. Comparative Cost-Effectiveness** How costly is this evaluand to consumers, funders, staff, and so forth, compared with alternative uses of the available resources that might feasibly have achieved outcomes of similar or greater value? Are the costs excessive, quite high, just acceptable, or very reasonable?	**10. Exportability** What elements of the evaluand (e.g., innovative design, approach) might make it potentially valuable or a significant contribution or advance in another setting?

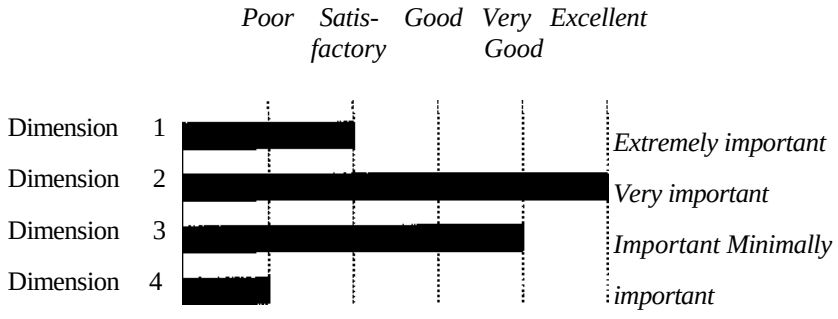
11. Overall Significance
Draw on all of the information in Checkpoints 6 through 10 to answer the main evaluation questions, including the following. What are the main areas where the evaluand is doing well, and where is it lacking? Is this the most cost-effective use of the available resources to address the identified needs without excessive adverse impact?

Determining the Importance of Dimensions or Criteria of Merit

Information about the importance of criteria can be used when profiling the performance of an evaluand on several different dimensions or criteria, as shown in Exhibit 7.2. Weak performance on a minor criterion (e.g., Dimension 4) may be no big deal, but weak performance on something really important (e.g., Dimension 1) would be very bad news indeed. Without this information

about importance, one might think that Dimension 4 represented the most pressing area for improvement or the evaluand's most serious weakness, when in reality, Dimension 1 should probably be the primary cause for concern.

Exhibit 7.2 Hypothetical Dimensional Profile With Dimension Importance Indicated



Determining the Importance of Evaluand Components

The same logic applies to component evaluation, where the evaluand is first broken down into components (or pieces), which are considered separately before looking at the overall picture. Both component evaluation and dimensional evaluation are analytical approaches and are distinguished from **holistic evaluation** (which involves considering the evaluand as a whole rather than breaking it down for analysis).

Component evaluation is common in the evaluation of policies, programs, or interventions that have several quite distinct parts. For example, suppose that a government policy is introduced with the aim of reducing juvenile delinquency. To achieve this goal, the government might implement several **policy instruments** (components or interventions) such as after-school programs for high-risk youth, more frequent police patrols in areas where juvenile delinquency is rife, counseling and guidance for first-time offenders, and tougher sentences for juvenile recidivists. When evaluating a multifaceted policy or program such as this, it makes sense to make the task more manageable by first breaking the evaluand out into components and considering each one separately before looking at the interactive effects and overall merit of the entire set of policy instruments.

Information about importance may be used when profiling the performance of the evaluand on each of its components (as shown in Exhibit 7.3) and/or when synthesizing the performances on multiple components to draw an overall conclusion about evaluand effectiveness (i.e., figuring out what all of the strengths and weaknesses add up to). Again, the information about \ importance allows us to (a) identify the components in most urgent need of improvement (if the evaluation is formative) and (b) have some basis for determining the overall merit of a package of interventions, some of which are working better than others (for drawing overall conclusions in a formative or summative evaluation). **(SEE EXHIBIT 7.3 PAGE 103)**

DETERMINING IMPORTANCE: SIX STRATEGIES

There are basically six strategies available for determining the importance of evaluative criteria or components:

1. Having stakeholders or consumers "vote" on importance
2. Drawing on the knowledge of selected stakeholders
3. Using evidence from the literature
4. Using specialist judgment
5. Using evidence from the needs and values assessments
6. Using program theory and evidence of causal linkages

Each of these strategies has advantages and disadvantages that make it a better choice in certain situations than in others. In this section, we look at how each one works and when to use it. **[PAGES 105-125]**

Strengths and Weaknesses of the Six Strategies

The six importance determination strategies outlined in this chapter vary considerably in their complexity and in the kinds of situations to which they are most applicable. Each has its own set of advantages and challenges, as outlined in Table 7.10. Often the best option is to employ the principles of critical multiplism (Shadish, 1994; Shadish, Cook, & Campbell, 2002) and to choose two or three complementary strategies with different weaknesses.

[TABLES at PAGES 126-127]

As noted in this chapter, one of the main uses of importance determination is to allow more illustrative profiling of findings for a client. However, there is another application: synthesizing mixed findings on several dimensions or components to draw an overall conclusion about evaluand quality or value. We discuss this challenging task later in Chapter 11. But first, we need to tackle another explicitly evaluative task: merit determination.

* EIGHT *

THE MERIT DETERMINATION STEP

Evaluation is the systematic determination of the quality, value, or importance of something (Scriven, 1991). That "something" can refer to an entire evaluand (e.g., a program or product), or it can refer to aspects (i.e., dimensions or criteria) or pieces (i.e., components) of an evaluand. The previous chapter outlined several strategies for determining the *importance* of evaluand components or dimensions. This chapter explores how to determine the *quality or value* of performance on these components or dimensions (i.e., the merit determination step).

> **Merit determination** is the process of setting "standards" (definitions of what performance should constitute "satisfactory," "good," etc.) and applying those standards to descriptive data to draw explicitly evaluative conclusions about performance on a particular dimension or component.

The merit determination step is where we apply the contents of the Values checkpoint to the descriptive data we gather to draw evaluative conclusions under the Sub-evaluations checkpoints of the Key Evaluation Checklist (KEC) (Exhibit 8.1). Note the explicitly evaluative questions under each checkpoint. The "big picture" question of how we should determine the quality, value, or importance of an evaluand overall is addressed later in Chapter 9, where we talk about synthesizing all of our findings to draw an overall evaluative conclusion.

DETERMINING MERIT: WHAT AND WHY

Merit determination involves two steps: (a) defining what constitutes poor, adequate, good, very good, and excellent performances on a particular dimension (or for a particular component) and (b) using that definition to convert empirical evidence of evaluand performance (descriptive facts) into evaluative conclusions (i.e., saying something explicit about quality or value). Here we are applying a basic evaluation formula:

Descriptive Facts About Performance	+	Quality or Value Determination Guide	=	Evaluative Conclusions

Merit Determination Using a Single Quantitative Measure

In the special case where performance is being measured on a single quantitative dimension, the quality or value determination guide would simply be a set of cutoff scores (e.g., for a test, > 90% = A/excellent, 80%-89% = B/good, 70%-79% = C/adequate). In some cases, it might be just one cutoff, that is, the line between satisfactory and unsatisfactory performance.

The difficult issue when converting scores to grades is determining where the cutoff score should be placed.

Exhibit 8.1 The KEC Checkpoints Where the Merit Determination Step Appears

6. Process Evaluation	**7. Outcome Evaluation**	**8&9. Comparative Cost-Effectiveness**	**10. Exportability**
How good, valuable, or efficient is the evaluand's content (design) and implementation (delivery)?	How good or valuable are the impacts (both intended and unintended) on immediate recipients and other impactees?	How costly is this evaluand to consumers, hinders, staff, and so forth, compared with alternative uses of the available resources that might feasibly have achieved outcomes of similar or greater value? Are the costs excessive, quite high, just acceptable, or very reasonable?	What elements of the evaluand (e.g., innovative design, approach) might make it potentially valuable or a significant contribution or advance in another setting?

Merit Determination With Qualitative or Multiple Measures

As mentioned previously, the use of a single measure to assess performance on a particular dimension is not generally good practice. This means that for most evaluations, the evaluation team will be faced with a much more complex set of data (often a mix of qualitative and quantitative data) that must be converted to evaluative conclusions.

Many of us are trained in either the social sciences or the hard sciences, so breaking things down into their component parts comes fairly naturally. Once we have done that, we can go out and gather the data while applying our knowledge of research methodology. The tricky part comes when all of those data come in and we are left with a mass of information that needs to be packed back together to find answers to the question of how well the evaluand did on a particular dimension or component (Exhibit 8.2)

A more concrete example of what the problem looks like is provided in Table 8.1. Here we have multiple sources of data pertaining to the performance of a hypothetical graduate program on just one dimension: job placement. To assess the program's performance on job placement, four **subdimensions** have been defined: (a) speed and ease of placement, (b) level and quality of jobs obtained, (c) prestige and desirability of organizations where graduates find work, and (d) match of positions with graduates' interests and aspirations. For each of the subdimensions, multiple measures and indicators have been collected.

USING RUBRICS FOR DETERMINING "ABSOLUTE" MERIT

So, how are we going to convert a mix of quantitative and qualitative data into some rating of the quality or value of that attribute or level of performance? One tool that can be incredibly useful (and a good conversation starter with the evaluation team and stakeholders) is a rubric.

A **rubric** is a tool that provides an evaluative description of what performance or quality "looks like" at each of two or more defined levels.

A **grading rubric** is a rubric that is used to determine *absolute* quality or value, whereas a **ranking rubric** is used for questions of *relative* quality or value.

(See Chapter 2 for a review of absolute versus relative quality or value.)

A generic example of a grading rubric that can provide a good starting point for the merit determination step is shown in Table 8.2. The rubric shown in the table is merely a starting point for rubric development. It does, of course, take a significant amount of additional work to define terms such as *exemplary performance* and *serious weakness.* This usually requires a combination of background research and extensive discussions with experts and/or key stakeholders. (Two simple examples of this are provided later in the chapter.)

Developing **Rubrics** in a Participatory Evaluation

This process of defining "how good is good" can be an incredibly valuable exercise for helping all sorts of organizations to think tiirough what they mean by *quality* or *value.* In a participatory evaluation, this part of die process forms an important part of the groundwork for the evaluation and doubles as an intervention that helps people to focus on what is really important about the work they do.

Whether the evaluation is being conducted in participatory mode or not, it is very important to talk to consumers at this point when developing a merit determination rubric. After all, the program, policy, or product is presumably designed to create value for them. This can help organizational staff to identify incorrect assumptions they might have been making about needs and other issues.

Table 8.2 Generic Rubric for Converting Descriptive Data Into "Absolute" (rather than "relative") Determinations of Merit

Rating	*Explanation*
Excellent	Clear example of exemplary performance or best practice in this domain; no weaknesses
Very good	Very good or excellent performance on virtually all aspects; strong overall but not exemplary; no weaknesses of any real consequence
Good	Reasonably good performance overall; might have a few slight weaknesses but nothing serious
Barely adequate	Fair performance; some serious (but nonfatal) weaknesses on a few aspects
Poor	Clear evidence of unsatisfactory functioning; serious weaknesses across the board or on crucial aspects

Sample Grading Rubric 1

To give an example of a more fully fleshed-out merit determination rubric, Table 8.3 shows what an early draft might look like for one of the subcriteria identified for a hypothetical master's program in evaluation. It is important to note that the example given in the table is merely a sample rubric that has not been subjected to discussion witii key stakeholders or job placement experts. In nearly all cases, rubrics such as this need considerable refinement based not only on, for example, student or graduate expectations but also on expert (e.g., recruiter, job placement specialist, employer) input regarding the job market and what expectations would be reasonable for graduates with this particular mix of qualifications and experience.

Recall the data collected for our hypothetical master's program:

- Three quarters (75%) of graduates who sought work found employment within 3 months of graduation (mean = 6 weeks).
- Nearly one third (30%) had job offers by the time they graduated.
- Only 15% were still unemployed, underemployed, or in jobs unrelated to their degrees 12 months after graduation.
- Most graduates (85%) complained mat finding work was considerably

more difficult than they had expected.

- The average graduate sent out 22 applications, was invited in for three or four interviews, and was offered one or two jobs.

Based on this information, the program in question seems to fit most closely with a rating of "good" on the rubric in Table 8.3. In a real evaluation, you might need to do some further digging to make sure that the rating is justified. For example, there might be a high proportion of graduates who have decided to go on to doctoral programs, in which case they should not be counted as having been unable to find full-time jobs.

Table 8.3 Rubric for Determining the Merit of a Master's Program in Evaluation on the Subcriterion "Speed or Ease of Job Placement"

Rating	*Description*
Excellent	All students had evaluation-relevant job offers on graduation or soon after (within 2 months excluding those who were not actively seeking such employment), and several students had more man one strong job offer. Several high-profile organizations recruited on campus or sought recommendations through program faculty to identify the best recruits.
Very good	The vast majority of students (> 80%) had evaluation-relevant job offers on graduation or soon after (within 2 months excluding those who were not actively seeking such employment), and several students had more than one strong job offer. A small number of high-profile organizations recruited on campus or sought recommendations through program faculty to identify the best recruits. Most students had to be quite proactive about networking and applying for jobs.
Good	Most students (> 70%) had job offers on graduation or soon after (within 2 months excluding those who were not actively seeking such employment), although some of these were not directly related to evaluation, and some students had more than one reasonable job offer. Most students had to drive their own job-seeking agendas quite hard, although some assistance was provided. Those students without job offers tended to be those who were more passive about job seeking.
Barely adequate	Most students (> 70%) had at least one job offer witiiin 3 or 4 months of graduation, although many of these were not related to evaluation. Job-seeking efforts had to be very intensive to obtain decent job offers. Many graduates reported that employers were not at all familiar with their university or the program.
Poor	With only a few exceptions (< 30%), most graduates of the program took up to 6 months to obtain placements (or promotions in their current jobs) that were only slightly better than what they had left to enroll in the program (or that were only slightly better than what bachelor's-level graduates were getting). Most positions were not related to evaluation but rather were related to the cognate areas.
Completely unacceptable	Graduates of the program found it difficult even to obtain positions equivalent to the ones they had left to enroll in the master's program.

Sample Grading Rubric 2

Here is another example of a grading rubric that draws on much more qualitative (i.e., nonnumerical) information to draw conclusions about merit. In this case, the rubric is used for performance appraisal; however, the logic is the same as for program evaluation. The rubric was designed for evaluating the performance of clerical staff in a small accounting office on their management of monthly accounts (one of several duties). It was developed in discussion with the two business owners/partners, who defined the expectations. The rubric starts with a description of the scope of duties, lists the main performance indicators, and then defines each level of performance—in this case, from "unacceptable" to "excellent" (Table 8.4).

Table 8.4 Performance Appraisal Rubric for a Specific Set of Tasks in a Small Accounting Firm

Monthly Support Packages (clerical)

Scope:

- Preparing monthly financial reports
- Responsible for data entry
- Liaising with clients
- Critically analyzing results
- Systematically reviewing income tax liabilities

Performance indicators:

- Timeliness and efficiency
- Accuracy
- Clarity of communication, tact, and diplomacy
- Use of Inland Revenue Department (IRD) or Internal Revenue Service (IRS) compliance knowledge

Instructions: Choose the description that best fits how well the objective has been met, and check the appropriate box.

Rating	*Description*
Totally unacceptable performance (score =1)	Any one or more of the following: (a) inadequate checking and/or following up of queries or missing information, leading to serious inaccuracies in data entry and/or monthly reports; (b) failed to report one or more major problems or issues to partners; (c) inadequate documentation, making auditing extremely difficult or impossible; (d) frequently rude or abrupt with clients; (e) failed to inform clients of important obligations on one or more occasions
Mediocre (substandard) performance (score = 2)	Any one or more of the following: (a) inadequate use of communication skills, checking or following up of queries, or missing information, leading to minor inaccuracies in data entry and/or monthly reports; (b) failed to report one or more minor problems or issues to partners; (c) failed to inform clients of minor obligations on one or more occasions, causing inconvenience; (d) barely adequate documentation and/or auditing trails, making quality checking possible but somewhat difficult; (e) inadequate prioritizing of time, leading to one or more jobs being completed outside budgeted time frames (except when delay was out of the accounting firm's control)
Good performance (expected level) (score = 3)	Efficient checking of data entry, allowing preparation of accurate monthly reports supported by clear work papers and audit trails; clients always informed of their obligations and requirements; partners kept informed of any problems or issues as they came to light; time prioritized so that all jobs were completed within budgeted time frames unless delays were out of the organization's control; queries and missing information always documented and followed up quickly and efficiently to ensure that jobs were not held up; all clients handled professionally and courteously with excellent communication skills displayed; thorough documentation, allowing for rapid evaluation of clients' overall financial positions and internal record keeping and systems
Performance exceeded expectations (score = 4)	All of the above in addition to the following: excellent use of communication skills and time management, ensuring that clients had an excellent understanding of their financial situations and that statements were 100% accurate and consistently completed well within budgeted time frames; meticulously organized work papers and audit trails, allowing any staff member to quickly ascertain the current state of any work in progress and to check the accuracy of work completed; constantly worked to streamline procedures for both clients and the accounting firm
All-around excellent performance (score = 5)	All of the above in addition to the following: superb professional service to clients, enhancing the reputation of the accounting firm and resulting in positive feedback and/or new clients through word-of-mouth advertising; innovative approach to managing monthly support packages, resulting in a smooth-running and error-free system that allowed jobs to be completed significantly more efficiently than time frames budgeted for (levels to be agreed on between partners and employees)

USING RUBRICS FOR DETERMINING "RELATIVE" MERIT

In some cases, the evaluation team will need to determine the relative merit (rather than the absolute merit) of performance on a particular dimension. Relative merit evaluations (i.e., ranking) tell us little or nothing about how good the performance was in any absolute sense. They simply tell us how the person or program did relative to peers or competitors, respectively.

"Grading on the Curve"

Perhaps the simplest example of this is the practice called "grading on the curve." Although the term *grading* is used (and letter grades may even be given), the instructor is actually ranking rather than grading evaluees. Table 8.5 shows a hypothetical rubric that might be used to generate grades for student performance in a large class.

The main problem with grading on the curve is that the letter grades imply that there is some sense of absolute merit (e.g., A = excellent, B = good, C = satisfactory). But the reality is that this system forces the instructor to fail a certain proportion of the class, whether those students are performing at an unsatisfactory level or not. In addition, it forces the instructor to give A's to 10% of the students, regardless of whether their performance was truly excellent. In general, if ranking is being used, the terminology used to label the categories should make it clear that this is ranking (e.g., "top 10%" instead of "A").

Standardized Tests

Most standardized tests, such as the Scholastic Aptitude Test (SAT), the Graduate Record Examination (GRE), and the Graduate Management Admission Test (GMAT), also determine relative merit rather than absolute merit, expressing scores in percentile terms that indicate the test taker's percentile rank (i.e., what proportion of all test takers scored lower). One recently added exception is the analytical writing section of the GRE, which provides a numerical rating that corresponds to a description of absolute merit. Tests of intelligence quotient (IQ) are another example of tests that determine where someone falls relative to the population. Unlike the aforementioned standardized tests, IQ

score ranges are assigned explicitly evaluative labels such as "gifted" (see Table 8.6 for the conversion rubric).

Table 8.5 Hypothetical Rubric for "Grading on the Curve" (actually ranking)

Score Falls in:	*Grade Assigned*
Top 10%	*A*
Next 25%	B
Next 50%	C
Next 15%	D
Bottom 5%	F

Relative Merit and Experimental and Quasi-Experimental Designs

Determination of the relative merit of outcomes is particularly important for experimental and quasi-experimental evaluation designs, that is, designs that incorporate the use of a control or comparison group. For example, student achievement scores for a particular school are often interpreted relative to state averages or by comparison with schools from areas with a similar demographic and socioeconomic makeup.

The usual approach of researchers using experimental and quasi-experimental designs is to assume that a statistically significant difference in the right direction is evidence of merit, whereas failing to attain statistical significance implies a nonmeritorious outcome. In evaluation, there is a need to look further than statistical significance—to practical significance.

A **statistically significant** result tells us only that any observed difference (or statistical relationship) is unlikely to be due to chance (e.g., a fluke sample yielding unusual data).

A **practically significant** result is one that translates to real impact on people's lives (e.g., the difference has a noticeable and nontrivial effect on functioning or performance).

When determining the merit of a particular outcome, it is important to take into consideration both its practical significance and its statistical significance (or the qualitative equivalent).

Table 8.6 Rubric for Interpreting IQ Scores

Evaluative Intelligence Rating	*IQ Score*	z *Score**	*Percentage Below*
Exceptionally gifted	160	+4	>99%
Highly gifted 7	145	+3	99%
Very superior/gifted	130	+2	98%
High average	115	+1	84%
Average	100	0	50%
Low average	85	-1	16%
Borderline	70	-2	2%
Mild mental retardation	55	-3	1%
Moderate mental retardation	40	-4	<1%

NOTE: a. A z score indicates how many standard deviations a score is above or below the mean.

SOURCES: www.psychologicaltesting.com/iqtest.htm and http://iq-test.learninginfo.org/iq04.htm

Using Comparisons to Determine Relative Merit

To determine the relative merit of a process, an outcome, or a cost criterion, it is important to identify useful comparisons. For example, the evaluation team might "benchmark" process, outcome, and cost criteria against what has been achieved elsewhere (e.g., by other evaluands of a similar scope).

Benchmarking is a systematic study of one or more other organizations' systems, processes, and outcomes to identify ideas for improving organizational effectiveness. It has been used in manufacturing for years and is now widely used throughout business and industry.

Benchmarking most commonly refers to a process of gathering comparison data about what organizations in similar or related industries are achieving. This approach to benchmarking focuses primarily on collecting quantitative data about process efficiency, outputs, outcomes, and costs.

Sometimes organizations undertake their own benchmarking studies. In such cases, two or more organizations (often doing business in different sectors) each agree to allow teams from the other organization(s) to come in and study their practices, compare results, and discuss how improvements were made. These studies are typically heavier on qualitative data gathering (e.g., observation of processes, interviews with key stakeholders), although they still look at specific quantitative data.

The following example, taken from an evaluation of a large-scale organizational change effort, illustrates the use of a simple rubric to determine the relative merit of evaluand components. In this case, the components were clusters of initiatives that formed part of the change effort. The rubric is designed to assess the relative cost-effectiveness of each cluster of change interventions (Table 8.7).

After drawing up the simple rubric, the next task was to rate each component (i.e., cluster of organizational change interventions) on the given scale. As an example, one of the components was a set of interventions intended to create a more strategic and constructive work environment. To this end, the organization had implemented a culture survey and a climate survey, that is, quantitative instruments sourced from separate providers that were to deliver periodic "snapshots" of the organizational culture and climate that managers would then reflect on before making changes in their business units.

Table 8.8 outlines the main costs and benefits of the two surveys and provides a list of alternative options given the resources the organization had at its disposal. By applying this information to the rubric in Table 8.7, this component of the organizational change effort was rated "below average."

Table 8.7 Rubric for Determining the Relative Merit of Organizational Change Interventions

Relative Merit	*Description*
Superior practice	Clearly the most cost-effective of the available alternatives
Above average	Considerably more cost-effective than most alternatives
Average	Approximately as cost-effective as most of the alternatives
Below average	Considerably less cost-effective than most alternatives
Inferior practice	Clearly the least cost-effective of the available alternatives

* NINE *

SYNTHESIS METHODOLOGY

Like merit determination, synthesis is another task that is very specific to evaluation. It is the tool that allows us to draw overall evaluative conclusions from multiple findings about a single evaluand.

Synthesis is defined as "the process of combining a set of ratings or performances on several components or dimensions into an overall rating" (Scriven, 1991, p. 342).

Synthesis is most relevant to the Overall Significance checkpoint in the Key Evaluation Checklist (KEC) (Exhibit 9.1). This is where the evaluation team needs to combine all of the evaluative information gleaned from looking at Checkpoints 6 through 10 (Process Evaluation, Outcome Evaluation, Comparative Cost-Effectiveness, and Exportability) to draw overall conclusions about the evaluand.

Exhibit 9.1 The KEC Checkpoints Where Synthesis Methodology Is Used

11. Overall Significance

Draw on all of the information in Checkpoints 6 through 10 to answer the main evaluation questions such as the following. What are the main areas where the evaluand is doing well, and where is it lacking? Is this the most cost-effective use of the available resources to address the identified needs without excessive adverse impact?

The form of synthesis covered in this chapter is not to be confused with meta-analysis or literature reviews. These involve summarizing or combining the findings of multiple research or evaluation studies (about different evaluands) to draw overall conclusions about the relationships among variables. Meta-analysis uses a very specific statistical technique to give a weighted average of effect sizes across multiple studies. As such, it can handle only quantitative studies. In contrast, a literature review uses the reviewer's judgment, rather than an explicit technique, to synthesize studies.

There is a substantial overlap between the merit determination and synthesis steps in an evaluation. Many readers likely noticed that the rubrics we used to combine a mix of data in the merit determination chapter are in fact a very simple synthesis methodology. In this chapter, we take that basic logic further with some more systematic methods that can handle more complex data.

SYNTHESIS: WHAT AND WHY

Nearly any evaluand has a range of strengths and weaknesses—some more important than others—that we need to consider when we draw evaluative conclusions about quality or value on a particular dimension or component or about the evaluand overall. After all, doing poorly on some aspect of minimal importance is less serious than doing poorly on something crucial. This is why we need synthesis methodology—to have a systematic way of taking into account the pluses and minuses uncovered when the evaluation team draws evaluative conclusions.