Evaluation Essentials

From A to Z

Marvin C. Alkin

EVALUATION ESSENTIALS

EVALUATION ESSENTIALS

From A to Z

MARVIN C. ALKIN

Anne T. Vo
Research and Editorial Assistant

THE GUILFORD PRESS
New York London

©2011 The Guilford Press
A Division of Guilford Publications, Inc.
72 Spring Street, New York, NY 10012
www.guilford.com

Printed in the United States of America

This book is printed on acid-free paper.

Last digit is print number: 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Alkin, Marvin C.
Evaluation essentials: from A to Z / Marvin C. Alkin.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-60623-898-1 (pbk. : alk. paper)
ISBN 978-1-60623-899-8 (hardcover : alk. paper)
1. Evaluation research (Social action programs) I. Title.
H62.A45675 2011
001.4—dc22
2010022239

Acknowledgments

First, my most sincere thanks to Anne Vo for her important assistance in preparing this book. Anne prepared early drafts of portions of several sections of the book and provided valuable research and editorial assistance at all stages of the writing. However, Anne's most important contribution was as a trusted colleague who was a thoughtful sounding board for all my ideas. Anne is among the very best doctoral students that I have had, and she will excel in the academic world.

Nicole Eisenberg, a former doctoral student of mine, is a superb evaluator currently working in the Seattle area. Her case study demonstrates a clear understanding of the context in which evaluators work and provides excellent fodder for discussion. Nicole is a wonderful writer and a clean thinker, and I am grateful for her contributions to this volume.

Several people read early versions of the whole book and offered helpful advice. An early version of the book was pilot tested on two people to obtain comments on content and readability. Rebecca Luskin, a doctoral student in my Social Research Methodology Division, was meticulous in her examination and offered a host of insightful comments. Allison Taka, an undergraduate student (and member of the UCLA women's basketball team), provided very helpful comments about understandability. I am also grateful to William E. Bickel, Ben Charvat, Chris Ethimiou, and Gary Galluzzo, who did blind reviews of the book and offered many helpful suggestions. Finally, my colleague

Tina Christie reviewed the "almost final" manuscript and provided wise and helpful advice.

A number of others offered substantive comments that most certainly improved the book. Eric Barela offered suggestions on the qualitative methods chapters. My UCLA colleagues Felipe Martinez and James Catterall made suggestions related to the sections on quantitative methods and cost analysis, respectively. Kyo Yamashiro, Tarek Azzam, and Annamarie Francois reviewed material as well. Assistance was also provided by several of my doctoral students: Lisa Dillman, Mark Hansen, and Jessica Robles.

As with each of my other books, it has been a delight working with my editor, C. Deborah Laughton. Deborah spent several years trying to persuade me to write this book and provided encouragement throughout the process. I am grateful for all of her assistance.

Last, but certainly not least, let me express deep gratitude and love to my wife, Marilyn. She had to live with me day by day as I obsessed with getting this book "right" (in my mind). Thank you, thank you, for putting up with me.

MARVIN C. ALKIN

To the Reader

HELLO

I think we should get acquainted, since we will be carrying on a conversation—a long conversation. I'm Marv Alkin. I've been a professor at UCLA for more than 40 years. Many years ago, I founded and directed the Center for the Study of Evaluation. Since then, I've written books on evaluation, done research on the field, and written many journal articles and chapters. But don't get me wrong—I'm not some ivory-tower type. My work has always been based on engaging in real evaluations and learning from doing. I probably have done more than 100 evaluations—mostly small- and middle-size scale.

I view evaluation skills as very cross-disciplinary. I have done many school program evaluations, both K–12 and higher education. I've also conducted evaluations of a psychiatric residence training program, a state's juvenile detention facilities, a self-actualization program for campesinos in Ecuador, an agricultural extension program in eight Caribbean countries, and many others. I love evaluation, and I'm happy that I fell into it. I hope you come to appreciate evaluation as well.

Now, a few comments about my personal life. I am married, with two children and six marvelous grandchildren (as determined by my unbiased evaluation). My avocational passion is college basketball, particularly UCLA basketball. I rarely miss a UCLA game, and go to some basketball practices as well. I never played on a basketball team, but when my son was very young, I coached his Junior Hi-Y team to two undefeated seasons—and then I retired from coaching.

Well, that's much too much about me. What about you?

WHO ARE YOU?

Actually, you are a number of different people. You might be a program administrator taking an introductory evaluation course (or a unit on evaluation) in your field. This course might be taught at the master's-degree level. This book is relevant to you because you use evaluations, you commission evaluations, and you are often engaged in ongoing evaluation of your program's development.

Perhaps, also, you are a member of a program staff reading this book at the suggestion of an evaluator. Some evaluators consider the most effective kind of evaluation to be one that obtains active participation from those who have a stake in the program. In reading this book, you might be able to participate more fully as a partner in the evaluation.

You might also be a beginning (or would-be) evaluator who is using this book in a first-course overview to the field. Furthermore, you might be using this book in a doctoral-level course as an accessible introduction to the field, supplemented by another text or by a variety of original source readings creatively selected by your instructor.

I welcome all of you to our conversation. For the first of you, reading this book will provide the eye-opening experience you desire. You will gain some understanding of evaluation and the processes involved. Your ability to potentially conduct evaluations will be enhanced by the opportunity to examine a case study (following Section B) and to apply newly acquired skills to that case example.

For the potential professional evaluators, this book is a start. You'll gain a foundation in evaluation, which can certainly be enhanced by examining the suggested further readings at the end of each section. You will, however, need other courses to expand on some of the technical aspects of evaluation.

And so join me, and let's have a talk.

Contents

Overview

Do you remember when in the fifth grade you were asked to learn all of the presidents and vice presidents (in order)? And then, most certainly, you were asked to memorize the state capitals. (Do you still remember them?) The question that I ask is whether these activities provided you with a real understanding about each of these states or about how government works.

In this book, I will not pepper you with the names of evaluation "capitals"—the names of evaluation theorists. After many years in the field as an evaluation researcher and theorist, I know the literature, and it is reflected in the writings of this book. Instead, I want to provide you with concepts—the ability to engage in evaluation. When people talk, when people converse, they don't stop after every second sentence and say something like "Jones, 2009." Let us converse about the process of evaluation so that you can "walk the walk" instead of just "talk the talk."

However, let me point out that some people might, at the conclusion of a conversation, express further interest in a topic and the desire to learn more. Thus, at the end of each section, I have provided some items for "further reading." Each of these suggested readings was selected because I felt that they were easily understood and not overly esoteric. Moreover, I generally have not recommended long articles or books. Finally, each "further reading" is accompanied by a statement consisting of a sentence or two indicating why I think it might be worthwhile to read.

Another means for further reinforcing evaluation understandings is provided by a case study scenario (the RUPAS case) to be found after Section B of this book. This case involves education, social welfare, community building, health, and so forth. It is potentially applicable to many fields. At the end of each of the subsequent sections there are questions to be answered or suggested group activities related to the RUPAS case. A group leader or instructor might further modify or adapt the case study questions to fit your field of study. (Note: Gaining Additional Understanding "Case Study Exercise/Resources/Further Reading" appear at the end of each section, starting with Section C. Only "Further Reading" suggestions appear at the end of Sections A and B.)

STRUCTURE

You might have guessed from the title that I am going to follow through with the "A-to-Z" theme. Yes, indeed. There are 26 sections in this book designed to teach you, sequentially, how to do an evaluation. I selected A–Z as a mnemonic device and as a way to break the sections into manageable pieces. However, let me point out that evaluation is not some mechanical, step-by-step valuing procedure. Furthermore, program site contingencies might alter the sequence and perhaps leave out steps. Evaluation involves people and interrelationships, and this is highlighted throughout the book.

Sections A and B provide some general understandings about evaluation: what is evaluation, why do evaluation. Section C is a "Who is the evaluator?" section. This is both general understanding and an important aspect in defining evaluation. The logic of this book is presented in the accompanying overview table. Then there are 13 evaluation activities roughly corresponding to Sections D through V, classified as to when they take place in the evaluation. Some commence primarily during an early (or "pre-") stage; others in what I call a "getting started" stage; some depict the completion of a written evaluation plan; and, finally, some activities involve executing the plan.

The remaining five chapters are of three types. Sections E, W, and X take place *throughout* the evaluation and are the "aids to getting it done properly." In Section Y, I present cost analysis as an evaluation option. And, in Section Z, I discuss with you some potential avenues for further learning. Look the chart over carefully, and then let us proceed.

Overview Chart: Evaluation Essentials

Evaluation activity	Section in which it is discussed	**The evaluation plan stages** Preplanning stage	Getting started on the plan	Writing the plan down	Executing the plan
1. Identifying Stakeholders	Section D	Primary	✓	✓	✓
2. Gaining Understanding of the Organizational/Social/Political Context	Section F	Primary	✓	✓	✓
3. Describing the Program	Section G	Primary	✓	✓	✓
4. Understanding the Program	Section H		Primary	✓	✓
5. Developing Initial Evaluation Questions	Section I		Primary	✓	✓
6. Considering Possible Instrumentation	Section J Section K Section L Section M		Primary	✓	✓
7. Determining Evaluable Questions	Section N		Primary	✓	✓
8. Finalizing the Evaluation Plan (Design)	Section O Section P			Primary	✓
9. Determining Procedural Aspects of the Plan	Section Q			Primary	✓
10. Analyzing Data	Section R Section S			✓	Primary
11. Answering Evaluation Questions	Section T			✓	Primary
12. Reporting Evaluation Results	Section U			✓	Primary
13. Helping Stakeholders to Use the Results	Section V	✓	✓	✓	Primary
Aids to getting it done properly					
Maintaining Relationships with Stakeholders	Section E	✓	✓	✓	✓
Managing the Evaluation	Section W		✓	Primary	Primary
Abiding by Appropriate Evaluation Standards	Section X	✓	✓	✓	✓
Additional evaluation option					
Conducting a Cost Analysis	Section Y	✓	✓	✓	✓

SECTION

A

What Is Evaluation?

Evaluation is taking place everywhere around us. You most certainly have engaged in evaluation within the past day. But what *is* evaluation? The popular definition of evaluation according to the dictionary is "to ascertain the value or amount of" or "to appraise." Indeed, you do this all the time. When you go to the store to make purchases, you determine the value of things. You might ask: Is it worth it? You look at the cost of an item and determine whether, in fact, its value to you exceeds the cost. You make an appraisal.

Perhaps the most common kind of evaluation that you might engage in is *product evaluation*. If you are looking to buy a new flat-screen television, you examine several different products. And when you do, you gather information about their technical specifications, their size, their attractiveness, and the cost. You make an appraisal. Sometimes these appraisals are done at an instinctive level. You might just look at competing products and make a decision, all while processing data in your head, perhaps unknowingly, about what you believe to be differences between the two products.

Sometimes, you or I might be more systematic in our evaluations. I recall that when my wife and I bought our first house, we listed the attributes that we thought were essential. Some items we considered to be necessary and other items were viewed as optional, but preferred. All of these attributes were listed on a piece of paper, and we developed columns for each of the three competing houses and rated each of the characteristics. Then the "evaluation model" became somewhat more sophisticated. We indicated those dimensions that needed to be present in order for the house to be considered (e.g., three bedrooms). This was a "necessary, but not sufficient" list. We then further differentiated between the houses by addressing additional ways of appraising the data. Values or weightings were attached to each of the dimensions. We needed to decide which ones, for example, were more important and then provided a weight for each. We asked: What are the weightings for each—the relative importance? Was having an additional bathroom more important than whether the house was landscaped well? How much more important? If landscaping was weighted at "1," would an extra bathroom be a "2" or a "3"? Thus in a way we were doing an evaluation based on a number of criteria weighted differentially based on our view of their relative importance.

Evaluating products—like a house—is one kind of evaluation. You might also evaluate people—*personnel evaluation*. You could make judgments about whether you would like to develop a friendship with an individual or whether a particular painter or electrician seems trustworthy and dependable. If you are in a position where you supervise personnel who work for you, you are engaged in evaluation. Or you might need to decide which of several applicants should be hired for a position. Personnel evaluations, again, require an appraisal, or an evaluation, including making judgments about relative value. Sometimes these kinds of decisions are made based on impressions—just instinct. Other times, those making such decisions are more systematic in performing these evaluations.

A third kind of evaluation is *policy evaluation*. Policies are general directions for action without necessarily having a particular program or plan in mind. So again, at the everyday level of evaluation, one might be evaluating a potential policy decision of whether to go on a diet. No specific diet plan is necessarily in mind; thus it is a policy being evaluated—not a program. This policy evaluation might consider such questions as, what are the potential benefits from commencing this policy—this course of action? In doing this, you might consider what you know about the relationship between being overweight and in good health. You might ask, "Is following this course of action compatible with my lifestyle, and, if not, is that acceptable? And what are the costs to me either in dollars or in terms of modifications that I would need to make in my lifestyle if I were to pursue that course of action or policy?"

Another kind of evaluation is *program evaluation*. Before discussing program evaluation, it is important that I add a brief side note. In program evaluation, evaluators can gather data about personnel (teachers, caseworkers, students, clients, etc.), but the focus is not to make judgments about these individuals. Products might also be a part of the program that is being evaluated. Thus data might also be gathered about products, but the primary purpose is not evaluating the products. Rather, evaluators are interested in using this information collectively to better understand the program in which participants or products are involved.

Now let us consider the nature of program evaluation. Suppose that you wish to enroll your child in a preschool program and need to make a choice about which one to select. Let me make the example simpler by assuming that you have become convinced of the benefits of the Montessori preschool approach, but there are three schools within easy driving distance that all claim to be "Montessori." In doing this evaluation, you might visit the three schools and observe in the classrooms. You might look at the activities in which children are engaged. You might look at the number of adults per child. You might look at the number and type of manipulatives available. All of these are relevant things to be examined. But if you wish to be systematic, you should select the kinds of things that are typically a part of a Montessori program—that follow the Montessori philosophy. When you have compiled that list of elements or activities, you must consider the possible ways to see whether those things are actually taking place. That is, you want to evaluate whether a Montessori approach is *really* being implemented—whether the program is really operating in a Montessori way. Thus you would want to examine: Does multiage grouping take place? Are there work centers? Are areas of study interlinked? Do children have a 3-hour work period available? Another element to examine is whether the teacher is Montessori-trained.

To the extent possible, you also want to examine what results are being achieved. Are the children happy? Have they increased in maturity? Do they have a love of learning? What have they learned?

To summarize, I have talked about evaluating *products, personnel* (or individuals), *policy*, and *programs.* In this book I focus on *program evaluations*.

PROFESSIONAL PROGRAM EVALUATION

Now let me separate the examples given above, which are everyday evaluations, from what I will call *professional* evaluation. As you have seen, there is great variation in the way that everyday evaluation takes place. These informal, nonprofessional evaluations range from somewhat systematic (perhaps even—or almost—"professional") to almost instinctual. For example, listing criteria and weighting them for relative importance as in the evaluation of various houses discussed above was relatively systematic. At the other extreme of everyday evaluations are those that are almost instinctual—a decision based on "I just had a gut feeling."

To be "professional," evaluation must be conducted in a systematic way. In essence, it is an inquiry involving the gathering and assessment of information in a planned and methodical way. Some authors use the term "disciplined" to describe activities such as professional evaluation and other forms of inquiry that are conducted in a systematic way. In this sense, disciplined inquiry refers to engaging in a procedure that is objective and one in which others are able to easily discern the steps that were taken. Finally, in disciplined inquiry, findings or conclusions have credibility. The manner in which the study was conducted must be so complete that the recipient of the evaluation has little doubt that the results are meaningful. Disciplined inquiries must carefully set in place procedures to consider errors in reasoning, data collection, or analysis of data. Credibility is established by paying heed to these potential sources of error and eliminating them, or, at minimum, exploring what they are and how they might influence the findings.

EVALUATION AND RESEARCH

Both professional evaluation and "research" are forms of disciplined inquiry. How do they differ? Sometimes the two are virtually indistinguishable. This is particularly true when considering evaluations performed by those who consider evaluation as basically a kind of applied research. But many other evaluators take quite a different look and assume pragmatic positions about evaluation, which are closely associated with reflecting users' needs and respecting their input and viewpoints.

The main distinguishing characteristic between research and evaluation is that the former *seeks conclusions* and the latter *leads to decisions*. Research seeks to add to the body of knowledge (typically of or pertaining to a particular academic discipline). Implicit in the concept of "knowledge" is that it is applicable across settings, across geography, and across time. By this I mean that the findings seek to be applicable *to like programs* anywhere, and be as valid in a year (or two or three) as they are now. Evaluations, as I wish to describe them, address the here and now (this program at this time) and attempt to provide insights that might lead to program improvement decisions. Evaluations recognize that there may be differences between programs that even have the same name. These differences are largely attributable to context—that is, the people involved and the particular situation.

Another important distinction between research and evaluation is "who asks the questions." Evaluation seeks to answer questions posed by, and of importance to, a client. Generally, a researcher defines the question he seeks to answer; researchers seek conclusions that add to understandings about the knowledge base.

Let me explore with you how disciplined inquiry is applied to an evaluation situation. In the example given earlier in this chapter, I discussed the evaluation of the Montessori schools. In that situation, the data collected would need to be justified as relevant indicators of the program characteristics by carefully studying the Montessori philosophy and other writings to discern the program elements that must be present to justify a program being categorized as Montessori. The procedures for amassing data would need to be considered nonarbitrary; rather, they must be well defined. What precisely does a particular program characteristic look like? You will need to consider: How will I unambiguously know when I see it? That is, how would I know that multiage grouping is taking place? In essence, what characteristics should be present? Furthermore, the person(s) gathering the data should be considered free of initial bias (or at least those biases should be specified as part of the evaluation). A legitimate professional evaluator should not enter the process with a predisposition to saying one or another program is best. Also, the way in which data are analyzed should be reasonable, easy to follow, and free of error. It should be patently clear how pieces of information (data) were analyzed (i.e., added together, or in some other way compared), or otherwise refined into more meaningful descriptions of results. Finally, the findings should be justified solely by the data. Evaluations may not take a broad leap to conclusions beyond the specific findings of the study.

EVALUATION DEFINITION

For those in need of a formal definition, let me provide one. But I will be brief. Formal definitions, detailed descriptions, and ponderous writing are not in keeping with the focus of this volume. Rather, I prefer to explain by offering examples and by raising rhetorical questions that lead the reader (you) to think about evaluation.

So, here we go. Most simply stated, evaluators state that evaluation is *judging the merit or worth of an entity*. This, in fact, is a statement of the *goal of evaluation*. The goal is to "value" in a systematic way. This valuing consists of two aspects. As you have seen, a part of judging is the determination of the *merit*—the intrinsic value of the entity being studied. The dictionary describes merit as intrinsic rightness or goodness "apart from formalities, emotional considerations, and so forth." Intrinsic goodness! What does *intrinsic* mean when I am talking about a program? If a program does well—that is, what it is supposed to do—it has merit. But is it sufficiently meritorious to satisfy the needs of a particular context? Think of my house-buying example. If a house has a large, ultramodern bathroom, then as a bathroom it might be considered meritorious but not have greater worth to me as a house buyer.

As we see, there are also *extrinsic* aspects to be considered. While the program may be meritorious, we ask what is its *worth* within our context? Is the high merit exhibited valuable within the particular program's context? Thus we seek to value or evaluate by considering both merit and worth.

The above provides a definition of evaluation based on its *goal*. Note, however, that I have stated that evaluation, along with research, is a disciplined inquiry. Thus we need to consider the *process* for reaching the stage of being able to judge merit and worth. This process requires systematic, unbiased context-sensitive behavior. In a sense, then, the various sections that I present in this volume are the *process definition* of evaluation.

A CONFUSION OF TERMS

Now let me deal with some of the confusing terms associated with evaluation. *Assessment* is a term that is often used synonymously with *evaluation,* but it is different. Another term that we often hear is *appraisal.* A very brief clarification is in order. My interpretation is that each of these three involve valuing (judging merit or worth). *Evaluation* is the favored term when we talk of judging a program. *Assessment* is employed when one refers to the clients of a program. This is particularly true in the education field, where we are constantly confronted with things like state assessment tests and national assessment of education. In each of such cases we are assessing students. *Appraisal*, I believe, is more relevant when we talk about program staff. Think of teacher appraisal, for example. Summary: We *evaluate* programs; we *assess* client knowledge; and we *appraise* staff.

Another kind of term is *testing.* I view this as different in nature from the above. *Testing is the process used for giving tests.* Tests are instruments for gathering data. They do not, in and of themselves, include a valuing component. They may subsequently be given value and enable judgments to be made. Thus I consider testing as a means of assessing, appraising, or evaluating.

Enough said.

RECAP—SECTION A

What Is Evaluation?

- Research and Evaluation—“Disciplined” Inquiry
  - Research—conclusion oriented
  - Evaluation—decision oriented
- Professional Evaluation
  - Product evaluation
  - Personnel evaluation
  - Program evaluation
- Evaluation Goal—Judging Merit or Worth
- Evaluation Process—Read This Book
- Other Terms
  - Assessment
  - Appraisal
  - Testing

EVALUATION PURPOSES

Another issue: Evaluation writers tend to distinguish between what they call “formative” evaluation and “summative” evaluation. *Formative evaluation* generally takes place during the early stages of the implementation of a program. It is conducted in order to provide information for program improvement. This generally means that the evaluation information would indicate how things are going. The evaluation information, for example, would highlight problems related to whether program activities were being conducted—and being conducted in a proper manner. Formative evaluation might also provide some early indication about whether program outcomes—the goals of the program—are potentially achievable. Did some early testing of clients show that they were not making sufficient intended progress? Formative evaluation is generally conducted primarily to benefit in-house staff. That is, it is information for those who are conducting the program so that they can make improvements. Such improvements might refer to modifications to ensure that the original program plan is complied with or might suggest changes in the program as conceived. The latter type of formative evaluation is the situation where evaluation results are used beyond fidelity concerns to re-form (form anew) the program. Michael

Patton refers to this latter type of formative evaluation as "developmental evaluation." In his conception, the evaluator's engagement is more proactive than is typical in most formative evaluations.

Summative evaluation is information designed to serve decisions—usually major decisions. This might mean making a decision about whether the program has been successful. Thus the results of a summative evaluation might lead to decisions about whether to continue the program or abandon it. A summative evaluation might also lead to decisions about implementing the program more broadly: "We have tried it out and it works. Let's do it at three other sites." Summative evaluations, thus, are primarily conducted for those who will make decisions about the program. These people might be administrators within the organization that sponsored the program, or they may be individuals within an external funding agency that has supported the program.

Robert Stake, a noted evaluation writer, is reputed as having offered the following pithy distinction:

- When the cook tastes the soup, that's formative.
- When the guest tastes the soup, that's summative.

Let us examine that distinction further. When the cook tastes the soup, he wants to decide whether all the ingredients were there. He thinks, "Did I put enough onion in? Should I have put in more?" If so, then he might change the written recipe. But there is another aspect to formative evaluation, and that is asking the question: "Did it taste good?" The first of these deals with *process*—the characteristics of what is included in the soup (or in an evaluation, this might be the various program activities). The second of these is looking at *interim outcomes*. Were the results positive? (In a program evaluation, this might be the same as looking at whether the short-term outcomes of the program were being accomplished.)

Obviously, then, when the guest tastes the soup, the major question is: "Did he like it? Was it good?" (That is, did it have merit and worth?) On the face of it, this would seem like a summative decision. The cook will consider whether the guest likes the soup in order to determine whether to continue offering the soup as a menu item. But perhaps there is more to it than that. What if the cook meets with the guests—the customers at the restaurant—and asks them how they liked the soup. What if *they* say it needs a bit more salt? Apparently, we have reached some summary stage wherein the cook has determined that it is appropriate to serve to guests. But there still is a formative element to the process. The cook might taste it again and decide that maybe it does need more salt.

And so I now propose an ever-so-slightly different description of evaluation purposes. I personally believe that a great deal of formative evaluation takes place in practice and only very occasionally do we conduct summative evaluations. More frequently, however, we engage in evaluation exercises that I would call "summary formative evaluation." That is, there is a formative period that has taken place, but at some point it is summarized or concluded. In my example, the cook decided to serve the soup. In a program evaluation, we frequently have an end-of-year evaluation report. It may be sent to program sponsors (given to the guests), but it nonetheless will provide information for modifying or improving the program (the cook will add more salt than was called for in the original recipe).

Furthermore, each of these evaluation purposes has both process and outcome elements associated with their conduct. A program process occurs—activities take place and interactions happen—and there are outcomes. Sometimes these are short term, like the learning that takes place at the end of a unit. Sometimes these outcomes are longer term, like at the end of the year or the end of the program. Think of these as evaluation "types." Note further that evaluation may have different *purposes*: formative (of two kinds) and summative.

Table A.1 depicts these evaluation purposes and evaluation types. Study the table.

TABLE A.1. Evaluation: Purposes and Types

	Types and audience			
Purposes of evaluation	Process	Interim outcomes	End-of-evaluation outcomes	Audience
Formative implementation evaluation	×	×		Program staff
Summary formative evaluation	×	×	×	Program staff, stakeholders
Summative evaluation	×		×	External audience, stakeholders

I point out to you now that there is another kind of evaluation question. Evaluators are often asked to work with stakeholders in determining program needs—referred to as *needs assessment*. In carefully examining the program as it currently exists and obtaining the views of those involved, the evaluator seeks to determine whether there are things that are believed to be *necessary* (not simply wanted) for the proper functioning of the program. As a further part of the needs assessment, evaluators might seek to examine the potential relevance of possible changes.

Now let me continue with my cook–soup example. Let us suppose that instead of making a soup, the cook felt that there was a need to add soup to his menu both because it would attract customers and would add a large-profit item. In the needs assessment, the evaluator might: look at menus of other restaurants; survey customers about whether they think soup would be a welcome addition to the menu; consider whether ordering soup might detract from their current ordering of salads; and what kinds of soup might be most appealing. This is a needs assessment.

As we proceed through this book, I highlight the particular aspects related to conducting an evaluation. I address the reasons for doing evaluation. I talk about evaluators and their capabilities. I consider who might be the most critical audiences for the evaluation. I consider how important it is to understand the nature of the program to be evaluated. Finally, I discuss the actual procedures involved in conducting an evaluation. The procedures will primarily be applicable to the two purposes of formative evaluation, but also will be relevant to summative evaluation.

GAINING ADDITIONAL UNDERSTANDING

Further Reading

In this section, and all that follow, I suggest some relevant further reading. I do not simply provide references. Rather, I have attempted to select, where possible, readings that are direct, to the point, and informative. Also provided with each is a brief comment indicating why I believe that the reading might interest you.

Bickman, L. & Reich, S. (2005). Profession of evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 331–334). Thousand Oaks, CA: Sage.

Bickman and Reich provide an excellent overview of the nature of the evaluation profession.

Mathison, S. (2008). What is the difference between evaluation and research and why do we care? In N. Smith & P. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 183–196). New York: Guilford Press.

Sandra Mathison provides a thoughtful discussion of the differences between research and evaluation.

Patton, M. Q. (2010). *Developmental evaluation*. New York: Guilford Press.

Michael Patton describes developmental evaluation as either a preformative stage and/or one that is especially applicable to evaluation of programs that are dynamic and which keep developing and adapting. See Chapter 1.

Scriven, M. (1996). Types of evaluation and types of evaluator. *Evaluation Practice*, *17*(2), 151–161.

Michael Scriven coined the terms "formative" and "summative" evaluation and talks about these terms as well as a host of other topics.

SECTION B

Why Do Evaluations?

As shown in the previous section, individuals are constantly placed in positions where they *need to make decisions*. They constantly have to choose. Likewise, administrators are placed in the position where choices need to be made between competing programs or courses of action. Ideally, these choices are based on a determination of which alternative is likely to produce the greatest benefit. And so, for example, in a health program, stakeholders might want to know which program improved the patient's health the most. Or in an education program, the issue might be student learning. Clearly, these are not easy things to measure. There are many facets to good health; one needs to be clear about the dimensions considered when a judgment is to be made on whether good health has been attained. Likewise, student learning has many facets aside from competency in reading, mathematics, and language. Expectations are that students will develop in other ways as well. Is problem-solving ability part of the desired educational outcome? What about attitudes?

While I have discussed decisions related to competing programs or courses of action, not all decisions are comparative. In some instances, program administrators might simply want to gain evaluative information about the status of a single program. This might lead to questions such as: Is the program operating in the way that we had anticipated? Are there any apparent deficiencies? Are participants satisfied?

MAKING DECISIONS

How are these various decisions to be made between competing programs? Typically, as decision makers examine the alternatives, many of them have a hunch about which they think to be best. These hunches or guesses are based on prior experience and practical knowledge. This practical knowledge is some combination of their own personal beliefs, interests, and experiences related to situations, which are in some way comparable to the decision at hand. Researchers refer to this as *working knowledge*. I certainly do not dismiss working knowledge as an important component in making decisions. Relevant and associated experiences certainly are important in understanding potential decision choices. But trust in our own instincts based on working knowledge alone goes only so far. Studies done by social scientists have documented the weaknesses and flaws in relying too extensively on such knowledge.

Another kind of personal knowledge is an understanding of the local situation; that is, those who would make decisions are influenced by the context in which they operate. They know their programs (or believe they do), and this *context knowledge* finds its way into practical knowledge or mindset; they trust their own perceptions of the program setting—its operation, its strengths, and its weaknesses. Of course, their perceptions are likewise not infallible.

Furthermore, programs about which decisions are being made sit within a political context, which influences decisions. These *political contextual concerns* exert influence on how decisions are made.

Clearly, there is also a need for more systematic data (information) to be a part of this decision process. This is especially true these days, given the extent to which the demand for accountability has become so prevalent in our society. People need to be convinced that program decisions, once made, were based on sound data. Enter the need for *professional evaluation*—disciplined inquiry directed toward a particular program and the potential decisions that might be made about it.

ISSUES FOR PROFESSIONAL EVALUATION

What kinds of issues does professional evaluation pursue? In the discussion above, I talked about *making program comparisons*—that is, making a choice between several programs. Sometimes, as in the Montessori example provided earlier, program staff or administrators might seek to make a choice between two or more programs currently in operation—but still, it is a comparison. Choose one. In professional evaluation we seek to eliminate biasing effects—that is, things that would make the comparison unfair. Each program must be considered in a comparable way with comparable conditions. The professional evaluation associated with such decisions is called *comparative evaluation.* Other times, however, a new program is being implemented, and the issue is whether it is worthwhile to consider this new program. That is, is it better than the existing program? The comparison, then, is with a program already in operation. This too involves a comparison. The issue becomes, "Compared to what?"

Some program comparisons focus not only on outcomes, but also on "how" the results were attained. Which particular aspects of the program had the greatest impact in obtaining the particular outcomes? In these cases, one is seeking to answer a causal question. *Causality* is extremely difficult to determine. Imagine a tobacco cessation program that included taking a particular medication, meeting monthly with a counselor, and meeting weekly with group participants. How does one determine which of these is responsible for attaining the desired results, or alternatively, the relative contribution of each? Typically, evaluation information for these types of decisions requires carefully controlled *experiments*. That is, we must create control groups that are randomly selected (i.e., a participant has an equal chance of being selected for any of the groups), and the intervention that the two groups receive should be the same except for one of the program characteristics. Then we are able to attribute the differences in outcomes or achievements to that single characteristic—there is a causal relationship. These kinds of studies are called randomized controlled trials (RCTs). Frequently, random assignment is not possible or warranted. In those instances, evaluators can attempt to provide indications of causality by conducting *quasi-experimental designs*. One such example is the use of a carefully selected comparison population—individuals or groups intentionally selected to match the control population (i.e., the individuals in the program being evaluated). Quasi-experimental evaluations provide less of a guarantee that causality can be truly established. High-level quantitative methodologists have derived sophisticated statistical models that can approximate causal conditions of RCTs, but we will not discuss those here.

We pause to note that it is extremely difficult to conduct such evaluations (experimental or quasi-experimental) in small programs, local or other. The number of program participants might be too small to attain random selection for the program and its "control" or comparison group. Moreover, the close proximity of program participants and their ability to communicate with each other leads to questions about whether the program and its comparison maintain true differences. These are summative evaluation questions and are not the primary focus of this book. However, this is discussed to some extent in Section N.

Some decisions, however, are not based on comparisons but instead take place within a single program and the basis for a decision might be whether a *particular standard* has been met (e.g., "Have 80% of the clients ceased smoking?" or "Have 75% of the students at the school achieved at the specified level on the federal standards established by the No Child Left Behind Act?"). Professional evaluation is especially relevant for decisions related to the determination of whether a standard has been met. Working knowledge clearly does not suffice in providing an answer to such a question. Hunches about something like having achieved a particular standard lack adequate specificity.

Another kind of standard that is often used is based on the results determined by a "normal" population on standardized tests. This is explained more fully in Section K.

Yet another type of decision might address issues related to the *implementation* of a specific program. This more basic kind of question refers to whether the particular processes envisaged were *in fact* implemented as planned. (In this case, we are dealing with something analogous to the patient compliance issue—did patients actually take the medication twice each week?) Or, as another example, did students in the classroom actually receive the instruction on a particular topic? In this case, the decision might be whether the particular attributes of the program—the activities that went on in the program—were in fact the ones intended. As I noted in the prior section, formative evaluation might also be proactive by not only examining fidelity, but also by working with program staff in modifying programs.

At a somewhat more esoteric level, the evaluation might seek to understand the *logic* of why certain actions take place within the program and their relationship to the desired outcomes of the program. That is, did certain program actions lead to unanticipated results, either positive or negative? (More on the logic of programs in Section H.)

Sometimes we do evaluations not for the decisions that are to be made, nor the decisions that will accrete. The role of evaluation in these instances is subtler—more future-oriented. Some evaluators envisage a broader purpose for evaluation. Their view is akin to the Chinese proverb about the greatest form of charity. To wit, "Give a man a fish and he will eat for a day. Teach a man to fish and he will eat for a lifetime." In the case of evaluation, the meaning is that evaluators seek to provide those associated with the program with a better understanding of their program and an *increased capacity* to understand evaluation, and to the extent possible, incorporate this into their regular activities. To achieve this evaluative purpose, evaluators strongly engage participants in the conduct of the evaluation.

Not all decisions are necessarily made at the conclusion of an evaluation. Sometimes there are *deferred decisions*, or decisions not necessarily intended to be made at a proximal point in time. In such cases, evaluations can add to one's understanding of a program. We know that evaluation is only one input among many that play a role in decision making. Other factors are involved, including costs, political feasibility, stakeholder values, and prior knowledge and decisions. One major evaluation writer, Carol Weiss, uses the lovely term *decision accretion.* Decisions do not just happen from an evaluation; they grow, they develop. Evaluation properly done should be part of that accretion. An evaluation, thus, might not lead to an immediate action, but could contribute to a knowledge base that aids in a later decision about the particular program under study.

Why do evaluation? We do professional evaluation in order to allow better decisions to be made (currently or in the future), to add to an organization's ability to learn about its program, and to further an organization's capacity to continue to benefit from evaluation. We care about improving programs in these many ways because we are incrementalists, and we know that in a small way this will help to improve society.

RECAP—SECTION B

Issues Addressed by Professional Evaluation

- Making Program Comparisons
  - Determining causality
    - Randomized controlled trials
    - Quasi-experimental
- Looking at Outcomes of a Single Program
  - Meeting preset evaluation standards
  - Comparison to test norms
- Looking at Programs Formatively
  - Examining implementation fidelity
  - Helping programs to change (developmental evaluation)
  - Examining the program's logic
  - Building an organization's evaluation capacity
- Providing Information for Deferred Decisions