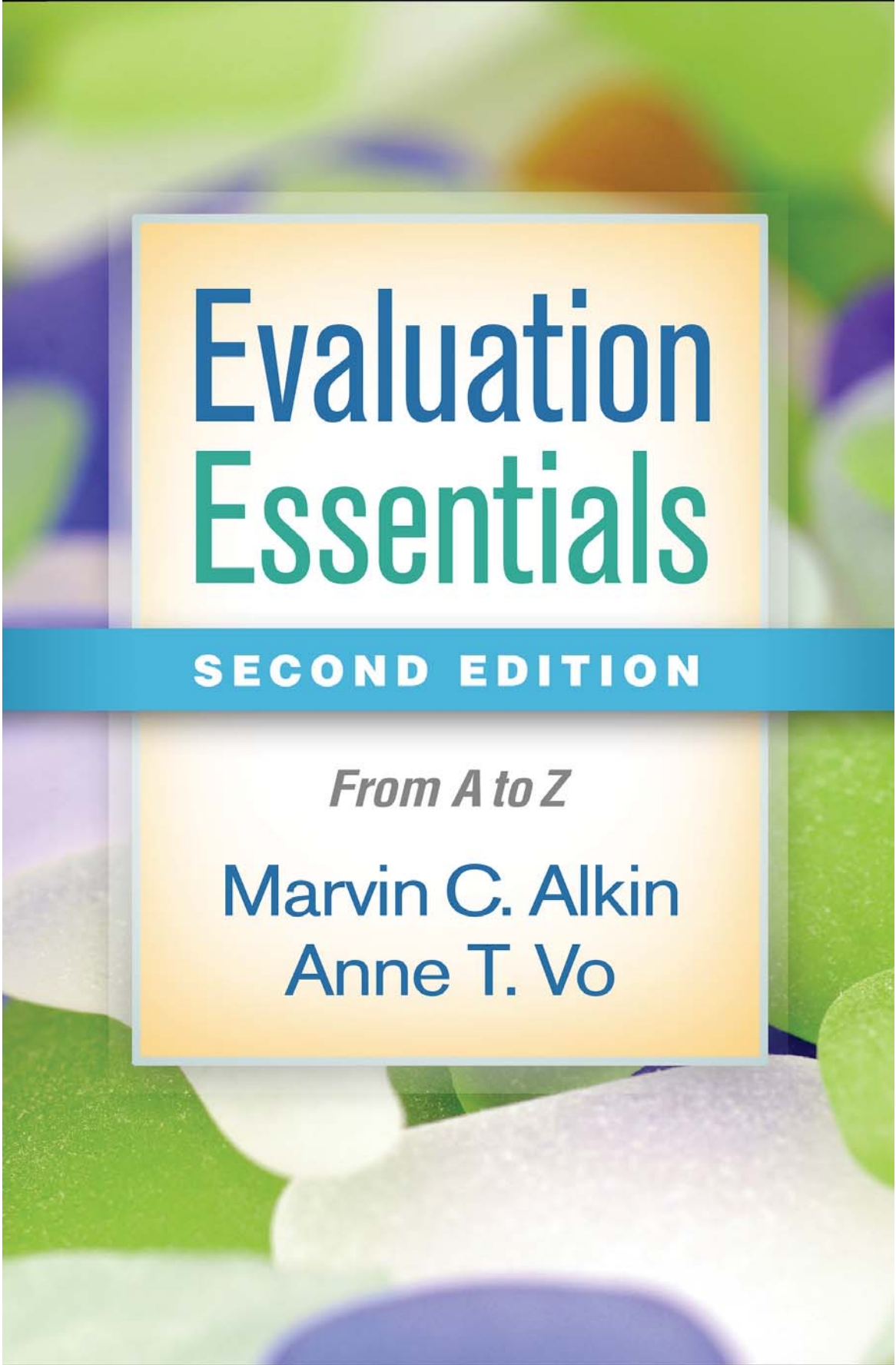
Evaluation Essentials
SECOND EDITION
From A to Z
Marvin C. Alkin
Anne T. Vo

EVALUATION ESSENTIALS

EVALUATION ESSENTIALS

From A to Z

Second Edition

MARVIN C. ALKIN
ANNE T. VO

THE GUILFORD PRESS
New York London

Copyright © 2018 The Guilford Press
A Division of Guilford Publications, Inc.
370 Seventh Avenue, Suite 1200, New York, NY 10001
www.guilford.com

All rights reserved

No part of this book may be reproduced, translated, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise, without written permission from the publisher.

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

Names: Alkin, Marvin C., author. | Vo, Anne T., author.
Title: Evaluation essentials from A to Z / Marvin C. Alkin, Anne T. Vo.
Description: Second edition. | New York : Guilford Press, [2018] | Includes bibliographical references and index.
Identifiers: LCCN 2017007304| ISBN 9781462532407 (pbk.) | ISBN 9781462532414 (hardcover)
Subjects: LCSH: Evaluation research (Social action programs)
Classification: LCC H62 .A45675 2018 | DDC 001.4—dc23
LC record available at *https://lccn.loc.gov/2017007304*

Acknowledgments

Evaluation Essentials: From A to Z was written by Marv Alkin in 2011. Motivation to write the book was rooted in a deep commitment to teaching evaluation, to supporting evaluation utilization, and to helping others to be active participants in the evaluation process. Anne Vo was a doctoral student at the time and prepared early drafts of several sections and provided valuable research and editorial assistance at all stages of the writing. Given Anne's early contribution as a trusted colleague and a thoughtful sounding board, Marv thought it only natural to renew this project with Anne, who is now a faculty member, as the book's coauthor.

This book could not have been possible without the generosity and support of many individuals. As with the first edition, Nicole Eisenberg, one of Marv's former doctoral students, provided an excellent case study for discussion. She is a wonderful writer and a clean thinker and we are grateful for her contribution.

We appreciate the very able clerical assistance that Kristen Gonzalez, Emi Fujita-Conrads, and Jacob Schreiber provided. We thank, as well, those who did blind reviews of the book and offered many helpful suggestions: Jill Ann Chouinard, School of Education, University of North Carolina at Greensboro; Robert J. Flynn, School of Psychology (Emeritus), University of Ottawa, Canada; Rick Sperling, Department of Psychology, St. Mary's University, Texas; and Carl Westine, Department of Education, University of West Georgia. Tarek Azzam's helpful

advice on issues related to data visualization has added tremendously to this volume. Also, Brad Cousins provided valuable insights related to evaluation theory. Finally, our colleague Tina Christie has been a source of wise and helpful advice.

It has been a delight working with our editor, C. Deborah Laughton. C. Deborah is the editor of six of Marv's books and has been thoroughly supportive throughout. We are grateful for all of her assistance. Also, our thanks to Nina Hnatov for her very able copy editing.

Last, but certainly not least, Marv wishes to express deep gratitude and love to his wife, Marilyn. In Marv's words, "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." Anne wishes to thank her parents, Nelson and Anita Vo, and her partner, Victor Toapanta, for their unconditional encouragement, patience, and support, all of which made this work possible.

To the Reader

HELLO

I think we should get acquainted because we will be carrying on a conversation—a long conversation. I'm Marv Alkin. I've been a professor at UCLA for over 50 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 have done over 100 evaluations—mostly small- and middle-size scale.

I view evaluation skills as 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 have 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.

Anne Vo is my coauthor. Anne is a former student of mine and assisted me in writing the first edition of this book. She is currently Assistant Professor of Clinical Medical Education at the University of Southern California and is responsible for teaching evaluation to physicians and leading a growing evaluation unit. In addition to studying how evaluations are done across different settings, Anne has lots of experience conducting evaluations. Her portfolio of projects consists of PreK–12 and graduate education programs, including those serving community college, law, medical, and other university students.

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

WHO ARE YOU?

Actually, you are a number of different people. You might be a program administrator taking an introduction-to-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 that the most effective kind of evaluation is one that obtains active participation of 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. We are confident that you will find this book to be very accessible and it will provide you with a firm foundation in evaluation.

You might be using this book in a doctoral-level course as an introduction to the field, supplemented by another text or by a variety of original source readings creatively selected by your instructor. For you, 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.

For 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 participate in an ongoing case study of an evaluation.

We welcome all of you to our conversation. And so, join us, and let's have a talk.

Contents

Note: Each chapter ends with a recap, thought questions, further reading, and quick read suggestions.

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, 2015."

Let us now converse about the process of evaluation so that you can "walk the walk" instead of just "talk the talk." While I will be the one talking with you, I want to point out that my colleague Anne Vo will be at our side. You won't hear her, per se, but know that she is with us.

While you will learn from our conversation, I want to 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 provide 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. In addition, I have included

some "quick reads," which are blogs from the American Evaluation Association that are usually only two to three pages long.

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.

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. Further, 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, respectively, provide some general understandings about evaluation: What is evaluation? Why do evaluations? These two sections are not included in the overview chart of evaluation activities (located at the end of this section). The logic of this book is presented in the accompanying overview chart. Section C is a "Who is the evaluator?" section. This is both a general understanding of what evaluators do, but more importantly a call to you, the evaluator, to understand who you are—to define your evaluation role. Then, there are 14 evaluation activities roughly corresponding to Sections C through V. Further, these activities are 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; others 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 F and W take place *throughout* the evaluation and are the "aids to getting it done properly." In Section X, I present cost analysis as an evaluation option. Sections Y and Z are not included in the overview chart. In Section Y, I present a theoretical model that I call "context-sensitive evaluation," which captures the essence of what we will have talked about in this book. 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 FUNDAMENTALS

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. Who Is the Evaluator?	Section C	Understanding who you are as an evaluator			
2. Contracting for the Evaluation	Section D	Primary	✓	✓	
3. Identifying Stakeholders	Section E	Primary	✓	✓	✓
4. Gaining Understanding of the Organizational/Social/Political Context	Section F	Primary	✓	✓	✓
5. Describing the Program	Section H	Primary	✓	✓	✓
6. Understanding the Program	Section I		Primary	✓	✓
7. Developing Initial Evaluation Questions	Section J		Primary	✓	✓
8. Considering Possible Instrumentation	Section K Section L Section M		Primary	✓	✓
9. Determining Evaluable Questions	Section N		Primary	✓	✓
10. Finalizing the Evaluation Plan (Design)	Section O Section P	(Primary)	(✓)	Primary	✓
11. Managing the Evaluation	Section Q	(Primary)	(✓)	Primary	✓
12. Analyzing Data	Section R Section S			✓	Primary
13. Answering Evaluation Questions	Section T			✓	Primary
14. Reporting Evaluation Results	Section U			✓	Primary
15. Helping Stakeholders to Use the Results	Section V	✓	✓	✓	Primary
Aids to getting it done properly					
Strengthening Relationships with Stakeholders	Section G	✓	✓	✓	✓
Abiding by Appropriate Evaluation Standards	Section W	✓	✓	✓	✓
Additional evaluation option					
Conducting a Cost Analysis	Section X	✓	✓	✓	Primary

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." 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.

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 to gather information about their technical specifications, size, attractiveness, and the cost. You make a valuation judgment. Sometimes these judgments are done at an instinctive level. You might just look at competing products and make a decision, all the while having processed data in your head, perhaps unknowingly, about what you believe to be the differences among the products.

Sometimes we 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 performed ratings with respect to each of the characteristics. Then, the "evaluation model" became somewhat more sophisticated. We indicated those

dimensions that needed to be present in order to be considered (e.g., three bedrooms). This was a "necessary, but not sufficient" list. We then further differentiated among 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. The question was 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.

A house is a product—evaluating products is one kind of evaluation. You might also evaluate people—a *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 make a decision about which of several applicants for a position should be hired. 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 these 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. For example, at the everyday level of evaluation, one might be evaluating a potential policy decision of whether to go on a diet. Because no specific diet plan is necessarily in mind, it is a policy being evaluated—not a program. This policy evaluation might consider things such 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 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. 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 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, but what do you look for? One approach is to focus on whether the scholars are truly adhering to the procedure established as being appropriate for and typical of Montessori schools—that is, are they truly implementing a Montessori approach? You might look at the activities in which children are engaged and the ratio of adults to children. 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—those that follow the Montessori philosophy. After you have compiled a list of elements or activities, you must consider the possible ways to see whether those things are actually taking place. In other words, you want to evaluate whether a Montessori approach is truly being implemented—whether the program is really operating in a Montessori way. Thus you would want to examine: Does multiaged grouping take place? Are there work centers? Are areas of study interlinked? Do children have a 3-hour work period available? Are the teachers Montessori trained?

You might also want to do an evaluation not focusing simply on the extent to which appropriate implementation had occurred, but on outcomes. The focus of an *outcome evaluation* is to obtain a good picture of the success of the program. Was the program successful? Is there a basis for a judgment that the program is worthwhile? Let us consider the kind of criteria you could use for making such a judgment. Perhaps there were expectations that students would have developed certain prereading skills. Possibly you were concerned about whether students would be better able to engage in socialization with their peers. Certainly, you anticipated that students would be better able to maintain appropriate behavior. While these and other outcome measures might be initially identified, the evaluation of the program should take heed

of other program accomplishments that you might not have thought about: Are the children happy? Have they increased in maturity? Do they have a love of learning?

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 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, the listing of 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. For these reasons, disciplined inquiry leads to findings or conclusions that 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 set in place procedures to carefully control potential errors in reasoning, and to ensure systematic data collection and 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 could 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 indistin-

guishable. This is particularly true when considering evaluations performed by those who consider evaluation basically as a kind of applied research, but many other evaluators (me included) view them to be quite different.

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 2 or 3) as they are now. Research seeks to gain generalizable knowledge. 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 could be differences among programs that might even have the same name—they could differ in goals or emphases or in their context—including the people involved and the particular situation. Evaluators respect these differences in context and seek to gain understandings only about the specific program being studied.

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, program, or community. Generally, researchers define the question they seek to answer and seek conclusions that add to understandings about the knowledge base.

Let me explore with you the way that disciplined inquiry is applied to an evaluation situation. In the example given earlier in this section, I discussed the evaluation of Montessori schools. In that situation, the data collected would need to be justified as relevant indicators of the characteristics of the program by carefully studying the Montessori philosophy and other writings to discern the program elements that must be present to categorize a program as Montessori. The procedures for amassing data would need to be considered nonarbitrary—that is, 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 pro-

gram 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 (meaning 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. Evaluators should not take a broad leap to conclusions beyond the specific data of the study.

EVALUATION DEFINITION

For those in need of a formal definition, let me provide one. 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 put, evaluators state that evaluation is *considering the merit and worth of an entity*. This, in fact, is a statement of the *goal of evaluation*. The goal is to consider value in a systematic way. This valuing consists of two aspects. As you have seen, a part of it is the determination of the *merit*—which is the intrinsic value of the entity being studied. The dictionary describes merit as intrinsic rightness or goodness "apart from formalities, emotional considerations, etc." Intrinsic goodness? What does intrinsic mean when I am talking about a program? If a program does well—that is, does 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, because "ultramodern bathrooms" are not something I value. A flat-screen television could have all kinds of fancy attachments. That television might have merit but if I do not value the attachments, it does not have "worth." Thus one must consider the extrinsic aspects of what is being evaluated. 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? We seek to value or evaluate through both merit and worth considerations.

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 constitute *my process definition* of what I call context-sensitive 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 brief clarification is in order. This is my interpretation. Each of these three terms 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 these 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 related term is "testing." I view testing as different in nature from the above. *Testing is the process used for giving tests.* Tests are instruments used for gathering data. They do not, in and of themselves, include a "considering," or a valuing component. They may subsequently be given value and enable judgments to be made. Thus I consider testing as a means of gathering the data for 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 make the distinction between what they call "formative" evaluation and "summative" evaluation. *Formative evaluation* generally takes place during the early stages of program implementation. Formative evaluation is conducted in order to provide information for program improvement, which generally means that the evaluation information would provide an indication of 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 it is likely that program outcomes—the goals of the program—are potentially achievable. Did some early testing of clients show that they were not making a sufficient level of 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 could suggest changes in the program as conceived. However, it might take place over extended periods of time; in such instances, this continuous formative evaluation is focused on the ongoing development of a program or innovation in more complex settings. Some theorists refer to this as "developmental evaluation."

Summative evaluation is information designed to serve decisions—usually major decisions. This might mean making a decision about whether the program has been successful in attaining its outcomes and documenting the final status of the program. Thus the results of a summative evaluation could lead to decisions about whether to continue the program or abandon it, or implementing the program more broadly: "We have tried it out and it works. Let's do it at three other sites." Summative evaluations, therefore, are primarily conducted for

those who will make major decisions about the program. These people could be administrators within the organization that sponsored the program or 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 do it right—did I forget to put something in that is part of my recipe?" If "no," he might then add the ingredient. Or perhaps an ingredient is of a different brand or coming from a different source. In that case, based on the formative evaluation, the recipe might require greater specificity about the source. Another aspect to formative evaluation 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.)

Now consider when the guest tastes the soup. Here the major questions are "Did he like it? Was it good?" (Or 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 the soup needs a bit more salt? Apparently, we have reached some summary stage wherein the cook has determined that the soup is appropriate to serve to guests, but there is still 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 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 stipulated in the original recipe).

Furthermore, as noted in the above example, 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 outcomes 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."

Table A.1 depicts these evaluation purposes and evaluation types. Study the table. While much of what is discussed in this book may be relevant to summative evaluation, the focus of this book is primarily on the two types of formative evaluation.

As we proceed through this book, I highlight the particular aspects related to conducting an evaluation and address the reasons for doing an evaluation. I talk about evaluators and their capabilities and consider who might be the most appropriate audiences for the evaluation. I consider how important it is to understand the nature of the program to be evaluated and its context and then I discuss the actual procedures involved in conducting an evaluation. The procedures are applicable to both formative and summary formative evaluation (and to some extent to summative evaluation).

TABLE A.1. Evaluation: Purposes and Types

	Types of evaluations			
Purposes of evaluation	Process	Interim outcomes	End-of-evaluation documentation	End-of-evaluation outcomes
Formative evaluation	×	×		
Summary formative evaluation	×	×		
Summative evaluation			×	×

GAINING ADDITIONAL UNDERSTANDING

Thought Questions

What typically comes to mind when you hear the word "evaluation"? How is the manner in which we have described evaluation in this section similar to or different from the ways in which you have experienced it?

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. Following each is a brief comment indicating why I believe that the reading might be of interest to you. In addition, I have included some "quick reads," which are blogs from the American Evaluation Association.

Mathison, S. (2008). What is the difference between evaluation and research and why do we care? In N. L. Smith & P. R. 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.

Stake, R. E. (2011). Program evaluation particularly responsive evaluation. *Journal of MultiDisciplinary Evaluation, 7*(15), 180–201.

Bob Stake describes different models that can be used to guide the conduct of an evaluation, with an eye toward a model that he developed and is refining—responsive evaluation—in this article.

Quick Reads

1. American Evaluation Association Guiding Principles for Evaluators
   http://tinyurl.com/ppkfslb
2. Susan Kistler on the AEA/CDC Summer Evaluation Institute Materials
   http://tinyurl.com/jnk23gk
3. Kirk Knestis on Innovation Research and Development (R&D) versus Program Evaluation
   http://tinyurl.com/hltv56q
4. Patricia Rogers on Ways of Framing the Difference between Research and Evaluation
   http://tinyurl.com/zjujgkd

SECTION

B

Why Do Evaluations?

As shown in the previous section, individuals are constantly placed in positions where they *need to make decisions* or gain understanding about a program. In fact, we constantly make choices in our everyday life. Likewise, administrators of programs and the communities they serve are placed in the position where choices need to be made—for example, between competing programs or courses of action. Hopefully, these choices are based on a determination of which of the alternatives is likely to produce the greatest benefit. For example, in a health program, stakeholders might want to know which program most improved the patient's health. Or in an education program, the issue could 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? What are the strengths of the program? Are participants satisfied? Are there any apparent deficiencies?

INFLUENCES ON INCREASED UNDERSTANDING AND DECISION MAKING

How are understandings about program status or potential comparative decisions made? Typically, as decision makers examine the alternatives, many of them have a hunch about which they think is 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 that 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 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 program context in which they operate. They know their programs (or believe they do) and this knowledge finds its way into practical knowledge or mind-set; they trust their own perceptions of the program setting—its operation, strengths, and weaknesses. Of course, their perceptions are likewise not infallible. Also, programs about which decisions are being made sit within a political context, which in turn influences decisions. These *political contextual concerns* exert influence on the way that decisions are made. We converse more about these kinds of influences in Section F.

Clearly, there is 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.

NEED FOR PROFESSIONAL EVALUATION

I have mentioned that the kinds of decisions to be made might pertain either to the status of a single program or its comparison with other programs. There are several ways that success may be determined within a single program. One kind of approach is 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 a particular federal act?") Professional evaluation is especially relevant for decisions related to the determination of whether a standard has been met—for determining whether the program is "accountable." 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. For example, when students are given a test we want to ascertain how they did compared with a population of students nationwide who also took this test—think of a test score and percentile on the SAT or GRE tests. (This is explained more fully in Section K.)

Yet another type of decision related to a single program might address issues related to the *implementation* of the program. This more basic kind of consideration refers to whether the particular processes envisaged were *in fact* implemented as planned. In this case, we are dealing with something analogous to the issue of patient compliance in medicine—did patients actually take the medication twice each week? Or another example—did students in the classroom actually receive instruction on a particular topic? In this case, the decision might be based on whether the particular attributes of the program—the activities that went on in the program—were in fact the ones intended.

In the discussion earlier, I talked about *making program comparisons*—that is, making a choice among several programs. Sometimes, as in the Montessori example provided in the first section, 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—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 becomes whether it is worthwhile to continue it—that is, is the new program better than the existing program? The comparison, then, is with a program already in operation. When making program comparisons, the question is typically, "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 weekly group participant meetings. 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 a control group that is randomly assigned (i.e., a participant has an equal chance of being selected for either group), 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, however, random assignment is not possible or warranted. In those instances, evaluators can attempt to provide indications of causality by conducting *quasi-experiments*. 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 do 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 one another leads to questions about whether the program and its comparison maintain true differences. These kinds of questions are not the primary focus of this book. They are useful if one is seeking to generalize to other contexts, but as we discussed in Section A, this is more of a research activity and not specifically appropriate for small-scale evaluations. (More on this in Section P.)

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, we might want to understand why it is thought that engaging in certain activities should reasonably lead to certain outcomes. In examining the logic, we may also discover program actions that lead to unanticipated results, either positive or negative. (More on the logic of programs in Section I.)

Sometimes we do evaluations not for the particular decisions that are to be made or for the decisions that might occur in the future. The role of evaluation in these instances is more subtle, more future oriented. Some evaluators envisage a broader purpose for evaluation.

Their view is akin to the old adage about the greatest form of charity. To wit, "Give a man a fish, and you feed him for a day. Teach a man to fish, and you feed him 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 evaluation capacity-building purpose, evaluators actively 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. Carol Weiss, a major evaluation writer, uses the lovely term "decision accretion." Decisions do not just happen from an evaluation—they grow and develop. Evaluation done properly should be part of that accretion. An evaluation, therefore, 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 to add to an organization's ability to learn about its program, to provide a basis for judging the accountability of programs, to allow better decisions to be made (currently or in the future), 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. We believe that a major evaluation concern for attaining use will foster these objectives.

RECAP—SECTION B

Why Do Evaluations?

- Influences on Decision Making
- Professional Evaluation in a Single Program
  - Meeting preset evaluation standards
    - Comparison to test norms
    - Examining implementation fidelity

- Making Program Comparisons
  - Determining causality
    - Randomized controlled trials
    - Quasi-experimental methods
- Other Evaluation Purposes
  - Examining the program's logic
  - Building an organization's evaluation capacity
  - Providing information for deferred decisions

GAINING ADDITIONAL UNDERSTANDING

Thought Questions

In your view, what purposes do evaluations serve? Have you seen evaluations being conducted for purposes other than what we have described in this section? If so, what are they and how has conducting evaluation for such purposes affected your organization (if you are a commissioner of evaluations) and/or your practice (if you are an evaluator)?

Further Reading

Cook, J. R. (2015). Using evaluation to effect social change: Looking through a community psychology lens. *American Journal of Evaluation, 36*(1), 107–117.

This article discusses how evaluations contribute to change processes in society.

Kennedy, M. M. (1983). Working knowledge. *Knowledge: Creation, Diffusion, Utilization, 5*(2), 193–211.

I consider this article a "classic." It describes quite completely the role of evaluation and other information in the decision process.

Weiss, C. H. (1980). Knowledge creep and decision accretion. *Knowledge: Creation, Diffusion, Utilization, 1*(3), 381–404.

Evaluation does not answer all questions immediately; answers to decisions "accrete." This classic article is very informative.

A footnote: I have always liked the Carol Weiss and Mary Kennedy papers. I strongly recommend that you read these articles. I know that decisions accrete and that working knowledge is a part of the decision process. Nonetheless, I believe that there are many instances where evaluation information has a primary immediate impact on decisions.

Quick Reads

1. Diane O. Dunet on Writing an Evaluation Purpose Statement
   http://tinyurl.com/h5nskph
2. Christine Johnson and Terri Anderson on the Quality Improvement–Evaluation Connection
   http://tinyurl.com/jz7w8or
3. Alice Walters on Appreciating Inquiry: Not Everyone Does!
   http://tinyurl.com/gwo8a58

Time-Out: The RUPAS Case

At this point, I ask you to take a time-out from your A-to-Z reading. You have learned about what evaluations are and why doing them has value. Before proceeding with the rest of the sections, I ask you to read the following case study. You will have chances to consider this case as you read the remaining sections. While the case is brief and does not offer the detail of an actual evaluation situation, you will have some opportunity to apply what you learn in each section of the book.

CASE STUDY

The Rural Parents' Support Program

NICOLE EISENBERG

THE PROGRAM

Family Matters (FM) was a community agency whose goal was to help disadvantaged families cope with the challenges of raising their children. For a little more than 10 years, FM had developed programs aimed at low-income families by providing information and social skills that could help parents support the healthy development of their children.

FM was located in a large city in the Pacific Northwest of the United States. Its headquarters were in an old area of town, in a rather run-down building where many of the other tenants were also nonprofit organizations. The interior of the agency's office was well kept, but crammed with desks, shelves, conference tables, computers, office supplies, and boxes upon boxes filled with paper.

The agency was run by Amy Wilson, a busy, vibrant, and cheerful woman in her late 50s. Amy and two colleagues—who were friends

Nicole Eisenberg, PhD, is a Research Scientist in the Social Development Research Group at the University of Washington. Her work has focused on the assessment of child well-being and on the evaluation of the Evidence2Success prevention model. Dr. Eisenberg has extensive experience evaluating social and educational programs for low-income populations in the United States and Chile, developing assessment tools and working with stakeholders to foster the utilization of evaluation findings. She has also done applied research in the areas of child development, early childhood education, and teacher effectiveness.

from their college years—had founded FM about a decade ago, and the agency had grown to a staff of 12 people. Aside from sharing a strong commitment to their work, the members of the agency had built friendships among themselves as well. Most of them were women who had university degrees in the social sciences (e.g., psychology, education, sociology, anthropology). Some had considerable experience in the field, while others were young and fresh out of college.

FM ran several programs, but their biggest one was the Rural Parents' Support (RUPAS) program. The goal of RUPAS was to provide training for parents living in remote, rural areas, so that they could better foster the development of their young children. The program, aimed at mothers for the most part, served families that lived in areas with little access to social and health services or early education centers. It basically offered parents training on how to stimulate their young children's development, in areas ranging from appropriate nutrition to cognitive and social skills. (RUPAS focused on the younger age group given that older children—6 years and older—attended elementary school.)

RUPAS was a social program with a strong nonformal education component. It used a parent education curriculum—developed by FM—that extended over a 2-year period. First, FM identified communities in need—basically, rural communities where there were no early childhood education resources, and where a number of families with young children resided. In each community, FM staff worked with community leaders (e.g., priests, school teachers or principals, pediatricians or family doctors) to identify potential mothers who might be willing to become "Parent Leaders." Parent Leaders would essentially become parent educators, who worked informally with groups of parents, passing on the information included in the RUPAS curriculum.

The mode of operation for RUPAS was that FM provided training to the Parent Leaders, who in turn offered workshops for other members of their communities. The training of the Parent Leaders took place in weekend-long workshops, where all of the Parent Leaders came together. Then each Parent Leader organized a group of parents in her community, and held weekly (or biweekly) meetings with them. FM supplied the Parent Leaders with materials that they in turn used in their group meetings. Parent Leaders received (1) a Parent Leader Booklet; (2) a Parent Booklet for each one of the parents (mothers) in the group; and (3) Child Booklets, with activities for the children of the participating families.

The Parent Leader Booklet contained information and suggested activities, with precise instructions for the Parent Leaders. The booklet was divided into chapters and subsections, and at each meeting parents focused on one of the sections. For example, there was a section on

nutrition with information about healthy diets for children, nutritional guidelines, suggestions for dealing with overweight (or underweight) children, recipes, and so on. There were also suggested activities for the parents, such as things they could do with their children or in their homes to promote healthy eating.

The Parent Booklets followed the same chapters and sections as the Parent Leader Booklets, and included the same information, except that they did not include tips specifically designed for the Parent Leaders, such as how to lead the group, what group exercises could help, how to approach certain hot topics, and so on. The Child Booklets similarly followed the same chapters and sections, but included activities specifically designed only for children (e.g., coloring pictures of fruits and vegetables with crayons). These child-focused booklets included activities for kids in a large age group (ages 0–6), but the parents and Parent Leaders received ideas on how to modify activities to serve children who were at different developmental stages and therefore had a diverse set of skills (e.g., simply read the story to the babies, ask toddlers to identify specific things).

The basic topics covered by the RUPAS program included physical health and nutrition, cognitive development, social development, emotional development, motor development, creativity, discipline and behavioral management, and school readiness and learning. The booklets used a conversational and friendly style, with straightforward language, no technical jargon, and plenty of graphics. The contents were "taught" in a very concrete manner, through numerous examples based on the lives of two fictitious children—Sally and John—whose experiences were common to these families.

FUNDING AND BUDGET

FM was funded in part by state funds, and in part by charitable foundations focusing on social and educational programs. FM first obtained seed money from the State of Washington to launch the RUPAS program. The funds were used, for the most part, to develop the curriculum, pay for salaries, and to train the staff. FM also secured extra funds from the Children's Trust, a local philanthropic foundation, and this money helped pay for travel costs involved in identifying appropriate sites for the program and for the publication of the educational booklets.

In terms of the program's budget, the first year was clearly the most expensive one, since the curriculum had to be developed and the staff had to be trained. After that, the program ran on a fairly low budget. FM staff trained Parent Leaders and provided ongoing supervision and

support as well as the educational materials: the booklets and other inexpensive supplies. There were costs involved with assessing needs in new communities and the obvious administrative costs, but overall, the program was quite efficient.

EARLY IMPLEMENTATION

For the staff members at FM, RUPAS meant a great deal. They devoted energy and expertise into creating their curriculum, deciding how to run the program, developing the booklets, and thinking about how to best approach their program participants. But despite their good intentions, the program had a rocky start. Like the old saying goes, "It is easier said than done." FM staff selected Rose as a key person to implement the program. Rose was in charge of identifying the communities, and then managing to find mothers willing to become Parent Leaders. She traveled to and from rural locations in Washington State to meet with people, and spent many hours talking with community leaders. Rose was an outgoing, friendly, and talkative soul, with a natural talent in establishing social relationships. During a short period, she had accumulated a large number of contacts and made important progress, but unfortunately, she had a personal family emergency that required her to take a prolonged leave of absence from her job. Given the unexpected nature of her family problem, Rose did not have time to explain to her new contacts what had happened and make a smooth transition out of her job. Finding someone to replace Rose and quickly reestablish her contacts proved harder than expected, especially since it was impossible to replicate the personal relationship that Rose had established with the families. This meant a big setback, since some community members felt "let down" by Rose's disappearance.

Eventually, most of the inconveniences were ironed out, and RUPAS began operating at six different sites. The sites were spread all over the state—some located more than 350 miles away from each other—requiring considerable travel for RUPAS staff.

COMMUNITIES AND FAMILIES

The communities where RUPAS operated had many similarities, despite their individual challenges and characteristics. They were all rural communities with a substantial number of families living under the poverty line. Although the communities varied in terms of their actual distance from urban centers, they all shared a lack of services and retail establishments.

The population served by the program was composed mostly of working-poor white families that lived in the countryside, either in old houses—some dilapidated and barely standing—or in tiny trailers. Many were two-parent families, but there were also plenty of single mothers. Most of the husbands worked minimum-wage service jobs—they served at restaurants or grocery stores, pumped gas, or did construction and roadwork—while most of the wives stayed home with the children. Some families worked in farm labor as well. Some parents did more than one job, because making ends meet for a family was just not possible on a single, minimum-wage income. The women who did work outside of the home had similar jobs, but unless they had family members or neighbors willing to provide childcare, it was tricky to find employment given the lack of day care for their children. There were also families on public assistance, especially single mothers caught in the dilemma of needing jobs but not having readily available or affordable childcare. Additionally, getting to and from work was a problem in itself. Living in the country meant being far from the jobs, and affording frequent vehicle maintenance in an area of dirt roads was often beyond their reach.

Some families had members who suffered substance abuse. Alcoholism was an old problem out in the country, but more recently, people were falling into methamphetamine abuse. Some people had started using drugs just to be able to keep up with crazy work hours and three jobs. In other cases, there were disruptive behaviors and mental health problems that had been undetected and untreated, probably in part because of the isolation and the lack of health and social services in the areas.

Most children in these communities stayed home with their mothers or relatives until they were old enough to go to kindergarten. When they were around 5 or 6 years old, they began to take the bus to school in the nearest town. At the schools, teachers often complained that the children were not "ready to learn" and lacked "basic literacy skills." Parents were glad to send their children to school, but sometimes did not understand the demands that the school required of them. There were few instances where teachers and parents could meet, offering very few opportunities to bridge the gap between them. Moreover, teachers felt they faced so many other problems—aside from unprepared children or distant parents—that they had little time to deal with these issues. The schools were underfunded and had difficulty recruiting qualified teachers, and students' scores on the state's standardized tests were low. Results from recent studies indicated that rural districts generally scored lower than their urban and suburban counterparts in the state. Approximately 30% of third graders met state standards in

the rural schools, compared with close to 41% in the urban ones and 57% in the suburban communities.

* * *

One of the areas where RUPAS operated—which in many ways was representative of other sites—was located in the northern part of the state. Although not the largest farming area in the state, this area nonetheless was a top producer of milk and berries (blueberries and red raspberries), and in 2007 its crop production had a market value of over $300 million.

Local people's lives here were paced—to a great extent—according to the seasons. Winter was long and wet, rain fell almost every day, and the sun was seldom out. Snow was less common but in some years, people got snowed in and spent weeks inside with nowhere to go. It was a beautiful area, with lots of green and water and mountains, but the lack of sunlight sometimes had devastating effects on people's moods, and there were few activities for young children when it was constantly cold and wet outside. Fall and spring were intensely beautiful and colorful seasons, and summer was probably the busiest for the farming families. It was also a period of increased tourism, which sometimes involved the benefit of extra cash.

Mary was the RUPAS Parent Leader in this community. She had two boys, who were ages 2 and 4 when she began the program. She was a stay-at-home mother, and her husband, Tim, worked at a big commercial farm. Tim sometimes spent many hours away from home, working extra hours at the farm to increase their income. This meant that Mary was alone with the boys for most of the day, and did all of the household chores. They lived in a trailer park a few miles from the farm. Their trailer home was cramped, and they had only two rooms. In the bedroom, she and her husband shared a bed, and the two boys shared another. The other room served as the kitchen, living, and dining room. They had an old dining set (a hand-me-down from Mary's sister when she moved out of town), a couch from the thrift shop, and a big, modern, expensive-looking TV. There were lots of plastic toys in a couple of bins and all over the floor, where the boys played. When possible, Mary bought some books for the boys, but the youngest one was quite good at tearing them apart—even the board books—so they did not last long in good shape.

Mary felt lucky that her husband had work and that they had a relatively safe place to live. However, they often had difficulty paying bills. Sometimes Tim brought food from the farm where he worked, but especially during the winter, when local fresh fruits and vegeta-

bles were not in season, the family's diet consisted heavily of pasta and canned soup. Mary tried to buy these in bulk when she and Tim traveled to the nearby town on weekends. But many times she ended up buying food at the overpriced market that was closest to her trailer. It bothered Mary that she could not provide for her kids as she would like to, but overall, she felt she was better off than many of the other people she knew, especially since so many people had been laid off from their jobs.

Mary decided to become a Parent Leader because she was eager to find activities for herself and her sons. She had always been an active person, but with the boys so young, it was often hard to get out and do things—not that she felt there was much to do around the area. She had been thinking about finding a job, but she really didn't have anyone to care for her boys. Before getting involved with RUPAS, she sometimes got bored just watching them play and doing housework. She loved her kids dearly, but sometimes resented how isolated she felt by being a stay-at-home mom. Her family lived far from her, so she had no one to help her watch her children. Like most parents, she struggled with figuring out how to best bring up her children, and the lack of role models had sometimes left her wondering what to do when faced with the issues of developing children: sleep problems, sibling rivalry, picky eating, tantrums, and so on. RUPAS seemed like a fun way to learn things that could be useful to her, as well as a great way to connect with other moms. She recruited a diverse group of parents and was a responsible leader.

RUPAS staff members were immediately happy with the decision to recruit Mary as a Parent Leader. They all liked Mary and admired her energy and good humor. However, they soon realized that Mary was unable to reach all of the families that RUPAS needed to serve, such as the families of the new Latino migrant workers who had recently joined the community.

CHALLENGES

Diverse Constituents

Latino families had begun to arrive in the state many years ago, mostly from Mexico and Central America, but the extent of their involvement in farmwork was larger now than it had ever been. In many parts of the state, Latinos were still a minority group, but in some towns in the eastern part of the state, they constituted the majority of the population. Some families had been in Washington for years and were legal residents, but other families were recent arrivals. Many of the people doing seasonal farmwork were undocumented Latino migrant work-

ers. Some of these laborers were men who left their families behind in their countries of origin, but others came with their families or had children once in the United States. These families differed from other working-class poor families mainly in that they did not have permanent homes or strong ties to communities outside of their families, which could be working across the state or across the country.

When offering program services, the staff at FM used a parent-education model that was based on weekly meetings that spread over a 2-year period—this system seemed to work well for families that FM was originally designed to serve. However, staff quickly learned that this model was not working for migrant workers who spent short periods of time in the communities, working the fields during a season, and then moving on to find other work when that was done. They were challenged to identify a service delivery model that was sensitive to mothers who were unable to commit to participating in RUPAS for a year or two.

Staff also realized not too long into the program delivery that they were not well equipped to support migrant families. Staff were largely monolingual and only a handful of program employees spoke a language other than English. Furthermore, the program materials were written entirely in English, which was useless in a community where Spanish or other home dialects was the dominant language. And because many mothers were undocumented, they were scared to participate in any program at all. They seemed to feel that the more invisible they stayed, the better it was for their families and their futures. But despite their invisibility, FM could not ignore them. The staff at RUPAS felt an obligation to reach out and serve them as well, but had trouble figuring out how to do it.

Program Implementation

At the same time, RUPAS staff members started noticing some new problems in the sites that they felt were implementing the program "successfully." There were many examples of situations that needed intervention.

On one site visit, Zoe—an FM staff member—met with a parent group in the eastern part of the state. Six mothers had met in a small room adjacent to a church, and were working on a chapter in the curriculum that addressed the topic of child discipline. The mothers began to talk about their own experiences as children, about being disciplined by their own parents, and the topic of spanking surfaced. This led to one mother disclosing being abused as a child. Her experience was traumatic and her feelings were raw. There was crying and suffering in the room. The Parent Leader did not know how to handle it. She felt

awkward, kept silent, lost direction of the meeting, and was unable to deal with the situation.

On a visit to another community, Zoe noticed that the mothers in the group were sitting around a table cutting out animal figures with scissors, and coloring the cutouts with crayons. The mothers were doing an activity that was intended for the children. The activity had detailed instructions in the Parent Booklet, but it was certainly not an activity designed for mothers to carry out. After the meeting, Zoe tactfully debriefed with the Parent Leader and realized that this was not the first time that a situation like this happened in this group. Somehow, the Parent Leader and the mothers had misunderstood which activities were meant for adults, and which were intended for the children.

Carmen, a Parent Leader, complained to Zoe that the mothers in her group spent the entire time chatting and eating cookies, and she didn't know how to get them back on track and keep the focus on following the program curriculum. The mothers had a great time, and showed up faithfully each week, but keeping the focus was difficult. On the other hand, Rebecca, also a Parent Leader, had the problem of nonattendance. The mothers were busy, their children very young, and although they liked the program, their turnout was always sporadic and irregular. When they did show up, they had missed so many meetings that they were "out of the loop" because of the content they had skipped. Rebecca had tried changing the schedule and the location of the meetings, but nothing was working. She was about to quit.

Thinking about the Future of the Program

Problems were not limited to what happened at the local rural sites where mothers implemented the program. Back at the community agency, Amy Wilson, the agency's director, worried about securing additional funding to continue with the program. The state's seed money covered a 4-year project but would eventually run out, and Amy wanted to secure at least another 2 years' worth of funding for RUPAS. They were now nearing the end of the second year, but she realized that given the time frames posed by the funding agencies, she would need to begin applying for new grants soon. She had confidence in herself and in her colleagues, and she had plenty of anecdotes to show how they were helping parents and children, but somehow these anecdotes did not seem enough to persuade other foundations and sources of funding. She also had questions of her own: "Have we made it clear what our expectations are?"; "Can we expand our focus to include the families of migrant workers?"; "How can we more efficiently manage the increasing costs of traveling to local sites?"; "Can the booklets and materials be improved?"; and "Are we meeting our goals?" She also

knew she had to find ways of demonstrating what was positive about the program and convincing possible funders that RUPAS was a worthwhile investment.

Amy devoted her time and energy into networking with different foundations. In doing so, she noticed that all of the potential new grants included evaluation requirements. She was surprised that over the years she had had so little exposure to evaluation. Just the word scared her. It reminded her of elementary school tests or doctors' physical exams. But it was something that was popping up more and more often in everything she read, and in every conversation with potential funding sources. She knew she had to think about evaluation and somehow include it in her work, but really didn't know where to start or how to go about it. She had her hands full just running the agency and program; it seemed beyond her reach to also have to "evaluate" it. But she wondered whether she really had a choice—it just seemed like something that *had* to be done.

* * *

DISCLAIMER

This case study is based on several actual programs that I have worked with over the years. However, it is a work of fiction that synthesizes aspects from different agencies, programs, and people—even different countries. Most of the "facts" presented are based on real situations that I have encountered in my work, but things that happened in one place are blended with those that happened in another, so that the result is an "imaginary tale" and not a reflection of any one real program.

SECTION

C

Who Does Evaluations?

Evaluations are done by people called "evaluators," but that's not all. Many people can and do conduct evaluation as long as they are engaged in the process of identifying questions, developing ways to answer them, and collecting and analyzing data in a systematic fashion. Clearly, larger and more formal evaluations are done by those who have the training to be identified as professional evaluators. Typically, this means that they have engaged in an academic program in which they have learned about the various steps involved in conducting an evaluation and have appropriate technical training in evaluation design, including data collection, data analysis, valuing, and reporting. Moreover, professional evaluators must have developed sensitivities about how to deal with stakeholders in order to appropriately understand the questions to be addressed and obtain their full participation.

But many other people in program settings engage in evaluation, or could engage in evaluation if they have the appropriate mindset toward systematically acquiring and using data to answer their program-relevant concerns. Evaluations can take place by either trained evaluators or others with some training and appropriate sensitivities to the particular setting.

EVALUATOR SETTINGS

One kind of setting for evaluators is when they are hired as outside consultants to conduct the evaluation. *External evaluators* (typically

well-trained professionals) are engaged to examine programs of which they are not a part. Hiring an external evaluator provides needed technical expertise and experience. Moreover, it provides the advantage of "distance"—that is, there is no presumed bias because the evaluator is not a part of the program nor formally involved in the program.

Sometimes, evaluators are *external* to a program, but *internal* to the organization encompassing the program. For example, these external–internal evaluators might be employed within a school district and have the assignment of evaluating a particular program in the school district. They are external to the specific program being evaluated in that they do not participate in it, but are internal to the larger organization. In this instance, the questions of distance and bias are mixed. On the one hand, these evaluators presumably do not have an interest in protecting the image of the program, but on the other hand, depending on the program's status and political potency within the organization and lack of total independence from the program, there is the possibility of evaluator bias. In this instance, as in the prior example of a fully external evaluator, evaluators are usually well-trained professionals.

Sometimes evaluators act in more of an *evaluation advisor* role, helping to facilitate the internal evaluation done by the staff of a program. The extent to which an evaluation advisor participates is quite variable. In one such instance, the internal staff may seek the advice of the consultant on some technical matters but would be responsible themselves for much of the evaluation. In other cases, the professional evaluator might be fully participating in the evaluation as an external evaluator but with a goal of full participation by stakeholders.

The team conducting the internal evaluation, as described above, could consist of an external–internal evaluator and an external advisor. This example might be taken to the next level where either internal staff appoint one of their own members as the evaluator for their program or the staff might function as a group and collectively engage in the evaluation. The results of these kinds of internal evaluations are often viewed as biased because the program's livelihood may impact the internal evaluator's continued employment. Thus the potential self-serving interest of members is often called into question in such cases as well.

For evaluations that are designed to form judgments about the quality of the program and to make decisions of major consequence about the program, internal evaluations are not recommended because of this potential deficit. However, where the purpose of the evaluation is to gain understanding of the program's processes in order to consider what has not been properly implemented and potential improvements in program activities, internal evaluation can be quite useful.

The intimate knowledge about the program by the internal evaluator or evaluation team simply adds to the evaluative understandings.

WHO ARE YOU?

I now ask you to consider who you are and how that might impact your conduct of the evaluation. Who are you? At the most elemental level you are defined by gender, ethnicity, and religion. These descriptors—individually and in combination—might influence how you think about (i.e., how you approach) an evaluation situation. In turn, it may impact how those involved in the program perceive you. Be aware of what this means for your role as evaluator and what, if anything, you can do about it.

Beyond these basic descriptors, there is a "you" who has beliefs, attitudes, and (yes) biases. Search your inner self to think about what these are because they could influence the way you think about the evaluation, the relationships you develop with program administrators and others, and the way you might do the evaluation.

There is also a professional side of "you"—the way you define evaluation and its purpose, which in turn is reflected in the way you choose to conduct the evaluation. Now let us take a few moments to consider several professional theoretical approaches that evaluators take in their work.

MULTIPLE ORIENTATIONS TO DOING EVALUATION

Who are evaluators? Do all evaluators go about their work in the same way? Do evaluators differ in the way they approach their work? Is there only one approach to doing an evaluation? There are many views about how to do an evaluation. These views are influenced by different conceptions of the purpose of evaluation—that is, those who write about the topic may have different feelings about why we do evaluation. Essentially, they differ on the primary purpose of evaluation. Let's explore some evaluation propositions (points of view) and I will tell you where I stand on each.

At its essence, evaluation is about assessing the merit or worth of an enterprise—a program. Does this program work? The reason for asking the question and the way in which one goes about finding the answer differentiates these different approaches to evaluation. In some

instances, one might want to determine the answer in order to make judgments about whether the program has complied with requirements. In my view, this is more of a "program monitoring for accountability" function, and not really evaluation. Another point of view seeks to determine whether the program achieved its intended end results—that is, evaluation is intended to assist in the making of decisions about the future of the program. (Yes, I agree with that, but there is more.) Another view of evaluation addresses a shorter time frame and seeks to gain understandings about the program, its processes, and its short-term accomplishments in order to provide information for program and organizational improvement and further development. (Yes, that makes sense to me.) Some evaluators view their work as having the potential not only to improve decisions about the program in question but also to provide insights for similar programs in other places (almost akin to research). (It's the "akin to research" that gets me. We learned in Section A that evaluation is concerned with *this* program at *this* time.) Furthermore, some evaluators feel strongly that evaluation has a moral obligation in the way in which it is conducted to specifically focus on promoting social justice within the program and more broadly. (Yes, I am for promoting social justice, but do not believe that ought to be the primary focus of the evaluation for improving this program now.) More recently, other evaluators have come to view evaluation as an agent for organizational change and learning—that is, evaluation is not just judging merit and worth, but through the engagement of the various participants, enabling them to better understand the process of acquiring information about their programs. (Yes, I agree with that.)

Now this may sound confusing. Clearly, evaluators may have different orientations to the way that they believe is the right course of action for conducting evaluations.

MY VIEW

In this book, I focus on what I call *context-sensitive evaluation*. The goal of this approach is the increased use of evaluation for improving programs—that is, I view the primary outcome of a well-conducted evaluation as fostering the use of the evaluation information for program change and improvement. I seek to ensure that the evaluation is not simply window dressing—that it has meaning and relevance and is, in fact, used. Evaluation use takes on many forms. I feel that the process of the evaluation should be oriented in ways that strongly engage

participants so that the likelihood of their subsequently using the evaluation results in decision making is enhanced. Use also extends not only to actual decisions but to gaining understanding of the program and changing views about it as well. I further recognize that the very act of program participants engaging in the evaluation is itself an important activity in terms of evaluation. This participation increases the potential for evaluation use by increasing the evaluation capability of the participants. The focus, then, is to possibly attain either short- or middle-term use of the evaluation findings in order to improve the program. The idea is that these insights will shed light on the program or contribute to the organization's improved evaluation capacities. In this way, meaningful changes of potential longer-term impact can be made.

The methods that I employ seek the active engagement—the participation—of a somewhat limited number of those directing and conducting the program, or who are recipients of the program, or who are otherwise influenced by it. I call these people a group of *primary potential users.* However, I seek to do so in ways that value the perspectives of broad stakeholders from diverse communities but without losing the virtues of intense participation by a smaller group of stakeholders.

I seek to adhere to the highest standards of methodological soundness without losing the focus on the local context and the specific program. My views are strongly grounded in my own research on evaluation use and the consideration of factors that are associated with a high level of evaluation use. These are characterized in four contexts: the user context, the evaluat*or* context, the evaluat*ion* context, and the organizational/community/political context. These form a total context for considering how to do evaluation, which I refer to as context-sensitive evaluation. I map this across five hallmark papers in Appendix C.

However, please be assured that all of the steps that I discuss in this book have applicability to other evaluation orientations as well. What may differ is the extent to which sections of the book have more or less salience.

Finally, it is important to note that while I have found it possible to adhere closely to a use-oriented context-sensitive evaluation role in most of the evaluations that I have conducted, there are instances where this was not possible. The context of the program and the expectation of stakeholders might have led me to the necessity of using a more hybrid form of evaluation—for example, in instances where there was a demand for findings that were more generalizable.

For you, as an evaluator, I urge you to be clear on what is expected and whether that role conforms with your interests and capabilities.

RECAP—SECTION C

Who Does Evaluations?

- Evaluator Settings
  - External evaluators
  - External to program, internal to organization
  - Evaluation advisor
  - Internal evaluations
- Multiple Orientations to Doing Evaluations
- My View—Context-Sensitive Evaluations
  - Use orientation
  - Deep stakeholder participation
  - Adapt as necessary

GAINING ADDITIONAL UNDERSTANDING

Thought Questions

My teaching philosophy has never centered on giving lectures. Instead, I seek to foster understanding through discussion and active learning. This book is an extension of that philosophy.

I have discussed evaluation in relatively broad terms thus far. Now I ask you to consider your previous evaluation experience. If you are an *evaluator,* how might you characterize your evaluation orientation? In what ways does your evaluation orientation influence the manner in which you go about evaluation? Consider also your personal views and attitudes and how they might impact your conduct of the evaluation. Likewise, if you are an evaluation *client* and are in a position to work with an external evaluator, in what ways could your awareness and understanding of the evaluator's orientation, ethnic background, or gender identity influence the nature of your collaboration?

Now that you have a general sense of these issues, I want you to consider the RUPAS program. We use RUPAS as a case example throughout this book to think through the evaluation topics, issues, and challenges that are discussed. In doing so, I hope to build on your existing knowledge and experience, that you will begin to get a feel for my evaluation "flavor," and that you have an opportunity to develop your own evaluation sensitivities.

Evaluation of RUPAS

With the RUPAS program in mind, begin thinking about how you might evaluate the program if contracted as the *external* evaluator. How would your positioning as an external evaluator influence decisions that you make along the way? In the

same vein, consider who might do an evaluation of that program internally. How would an *internal* evaluation be different? In what ways could internal and external evaluators collaborate?

If you are reading this book as part of a class, why not have a discussion about this with other readers?

Further Reading

Barela, E. (2015). Evaluation use and the internal evaluator: A balancing act. In C. A. Christie & A. T. Vo (Eds.), *Evaluation use and decision making in society: A tribute to Marvin C. Alkin* (pp. 31–52). Charlotte, NC: Information Age.

In this chapter, you will find a careful analysis of different kinds of tension that an internal evaluator must consider.

Barrington, G. (2005). External evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 151–152). Thousand Oaks, CA: SAGE.

The author provides a brief summary of the distinctions between internal and external evaluation.

Brandon, P. R., Smith, N. L., & Hwalek, M. (2011). Aspects of successful evaluation practice at an established private evaluation firm. *American Journal of Evaluation, 32*(2), 295–307.

This article offers a description of an exemplar internal evaluation firm and ways in which excellent practice is supported therein.

Volkov, B. B. (2011). Beyond being an evaluator: The multiplicity of roles of the internal evaluator. *New Directions for Evaluation, 132,* 25–42.

The various roles that an internal evaluator plays is examined in this article.

Quick Reads

1. Eun Kyeng Baek and Seria Shia Chatters on the Risks in Internal Evaluation
   http://tinyurl.com/zov75zx
2. Charmagne Campbell-Patton on Considerations for Evaluation Use: Interval versus External Evaluators
   http://tinyurl.com/hdjen6b
3. Erica Roberts on Organizational Culture and Internal Evaluation
   http://tinyurl.com/h53rwh8
4. Fred Seamon on Evaluation Careers in the Private Sector
   http://tinyurl.com/hdwkl3t

SECTION

D

Contracting for Evaluations

So how does one get to do an evaluation? The stimulus for the conduct of a program evaluation usually comes from a perceived need within the program. It might be that those administering the program feel that the goals that they had set for the program are not being accomplished. Sometimes an organization has applied to a governmental entity or other fund-granting agency for money to modify their program or to create a new one. Frequently, these agencies require that an evaluation be performed so that they might gain an understanding of whether their funds were well spent—that they had accomplished the intended goals.

In this section, we examine the activities that you, as an evaluator, might become engaged in when you seek to obtain a contract for the provision of evaluation. Contracting activities will be discussed in four parts, which are labeled Acquiring an Evaluation, Writing the Proposal, Preparing the Contract/Agreement, and Developing the Budget.

ACQUIRING AN EVALUATION

Evaluations are commissioned in a variety of ways. Some evaluations are conducted within the organization itself. These can be categorized as *internal* evaluations. In such instances, someone within the organi-

zation is asked to lead the evaluation of the program—one with which he or she is familiar. Hopefully, that individual has some evaluation expertise. Perhaps that person is you, and I hope that this book is helpful in performing that assignment. Where the individual who will be performing the evaluation is a member of the *program staff,* there may very well be fewer written documents—that is, there might not be a need for a formal contract, but there should at least be an agreement stipulating what is to be done and perhaps the additional resources that will be needed for conducting the evaluation.

Sometimes internal evaluations are conducted by individuals not specifically within the program to be evaluated, but from the larger organization that encompasses the program. We referred to this as *internal–external* in the previous section. In this case, the resources for conducting the evaluation might already be within the evaluation unit. In some instances, supplemental resources might be provided by the program itself. The nature of the relationship and the necessity for formality in agreements will differ from case to case.

Most evaluations, however, are called *external*—that is, they are conducted by a party outside the organization who has been specifically commissioned for this activity. This is where I will focus most of my comments with respect to the process of acquiring the contract or agreement to perform the evaluation. Contracting for external evaluations takes on several forms. Sometimes, program personnel are aware of an evaluator whom they believe to be competent and will approach him or her about potentially conducting the evaluation. In the evaluation literature, this is referred to as a *sole-source evaluation*—that is, only one person is contacted. In more instances than not, the evaluation report commissioners have a fixed dollar amount in mind and are simply asking the potential selected evaluator to indicate what might be done in the evaluation within that particular cost constraint. Sole-source evaluation normally requires a relatively simple agreement that I will shortly describe.

More typically a program will allow *multiple potential evaluators* to compete in acquiring the evaluation. This might be done informally by simply contacting a number of individuals or agencies that are known to conduct an evaluation of the particular kind of program in question and providing them with a general understanding of what is required and the resources that are available. Sometimes this selection from multiple evaluators is done at a more formal level. In that instance, what is referred to as a *request for proposal* (RFP) is employed and distributed broadly so that individuals can decide whether they want to be considered in the competitive process.

WRITING THE PROPOSAL

Typically, several factors are considered in the selection of an external evaluator based on the proposal. To put it simply, we might consider the three components as the *evaluation objectives and plan,* the *budget,* and the *competency* of the evaluators. There are great variations in the ways in which these components are presented and are emphasized. Some RFPs provide fairly elaborate descriptions of the evaluation activities and design that is desired and seek cost estimates from competitors along with an indication of staff competency.

In the second, most prevalent alternative format, RFPs provide a maximum dollar bid and a minimally developed plan, asking the potential evaluator to provide a detailed indication of how the evaluation might be conducted. In this instance, the evaluation bidder must intuit the principal stakeholders' important questions, as well as the logic of their program activities. Based on this limited information, the proposer would present a design for evaluating the program. In such a case the evaluator, guided by initial directions presented in the RFP, creates the evaluation design based on his or her own views as to what might be appropriate. Sometimes this initial proposal is quite fixed and the intent is that it will be stringently adhered to. Other times there is flexibility for minor changes based on subsequent discussions with program managers or other users.

Clearly the proposal must stipulate the individuals to be involved in conducting the evaluation and should include information about their capability, experience, and competence related to conducting an evaluation in this particular type of context. Moreover, there should be an indication of the time commitment of each of the key professional personnel. In simple agreements, this might be a rough estimate of the percentage of time to be devoted to the evaluation effort. In more complex proposals, time might be stated as the number of days of effort or, even further, expressed in terms of time to be devoted to each of the designated tasks required for accomplishing the evaluation. The proposal would also provide a budget for the evaluation. I discuss this later in this section.

In my view, the most ideal alternative is when the proposal presents broad guidelines and a plan for how the evaluator and users will interact to gain understandings that lead to a feasible evaluation plan that is directed toward user concerns and would be considered relevant to their program needs. As you will discover in the remainder of this book, program stakeholders often do not explicitly articulate the logic of their program (see Section I). By this I am referring to the reasons why they are doing particular activities and what they hope each will

contribute toward the accomplishment of program goals. Frequently, there is not a clear understanding about the evaluation questions that are of greatest importance (see Sections J and N), which is why I discuss the development of the evaluation design and the plan for enacting and managing the evaluation later in Sections O through Q. An evaluator in the situations of the first two alternatives presented above may need to consult these sections of the book for guidance in the development of the evaluation design and plan for his or her proposal. Although these issues will be addressed in an initial detailed proposal, hopefully there will be opportunity to revisit the topics when more is known about the program.

PREPARING THE CONTRACT/AGREEMENT

Some evaluation writers are absolutely adamant that every evaluation should have a formal contract with all of the "i"s dotted and all of the "t"s crossed, so to speak. This is certainly sound advice if the evaluation is large enough. Large evaluations with costly budgets, many tasks, and numerous personnel deserve to have the understandings between evaluators and contractors well stipulated—this is certainly the case for contracts from governmental agencies and from most philanthropic foundations. Indeed, these agencies usually have specific contract requirements. Unfortunately, if a project is small, a relatively substantial amount of energy, evaluator time, and implicit financial costs may be expended in simply negotiating a complex contract. It simply makes no sense.

However, there must always be a "meeting of the minds" between those who contract for the evaluation and the evaluators. Some kind of written agreement is always necessary—this is helpful in avoiding potential future conflicts. For very small evaluations, a letter from the evaluator stipulating the various terms and understandings might be sufficient. This certainly would include a statement of the contract fee, payment schedule, and responsibilities of the program. Also included would be understandings about the potential areas of interest to be evaluated and some critical thoughts about how the evaluator might pursue the inquiry. It is helpful, also, if the evaluator indicates the actions to be taken to better understand the program and its evaluation needs. This evaluator proposal letter should be accompanied by a confirmation letter from the contracting agency agreeing to those terms. The degree of completeness and formality will vary based on the size of the program (and of the evaluation).

But if there is a more formal contract, what needs to be stated? What needs to be agreed to? What is the essence of a contract or even

of an informal agreement? If there was a written formal proposal, then the technical details of the design, budget, and management aspects that were a part of it should be included in the contract.

As noted earlier in this section, contracts and agreements may be reached and agreed upon at various points in time. Assume for a moment that the contract (and we will use this term to mean both formal and more informal agreements) had been signed based on a specified evaluation plan. In that instance, you, as the evaluator, would not have had the opportunity to gain full understanding of the program, its logic, and the development of finalized questions to be addressed. That being the case, the evaluation contract should specify the basis for negotiating modifications in the plan based on a newly acquired understanding about the nature of the program to be evaluated. The contract should indicate the initial understandings about the goals of the evaluation, the initial preferences about the kinds of data to be collected, and (to the extent possible) the kinds of analysis procedures to be employed. But let me note here that if you, as the evaluator, had provided a plan upon getting the contract, it might not have been as detailed as what is described in Sections O through Q. You may not have had the same level of familiarity with the program. Under those circumstances, the plan as described in the contract might need to undergo changes as you help the primary stakeholders to define their questions and examine potential evaluability. Such changes, if made, should culminate in a written agreement about the revisions made.

The contract should also address the issue of evaluator reporting responsibilities. Will there be a final evaluation report? Some evaluators maintain that there is not always a necessity for a final written evaluation report. I believe that written evaluation reports are usually necessary. Thus, it is important that the evaluator reach an agreement about the reporting time line with the stakeholders. When is the final report due? Are there to be other interim written reports as well and on what schedule? If oral reports are to be provided, there should be agreement on when that is to occur. An allied issue is a designation of the responsibilities of the program and the primary stakeholders. Included here might be such things as the provision of on-site office space, providing access to staff and program recipients, and the provision of particular extant data and reports. Aside from these issues there are other evaluator responsibilities that should be noted in an agreement or contract. One major such item is a note that the evaluator will abide by appropriate ethical and technical standards (see Section W). Also included in the contract are such things as keeping data secure and maintaining confidentiality.

DEVELOPING THE BUDGET

The budget constitutes an agreement with the client on an *established dollar amount* that will be made available for conducting the evaluation. Sometimes the dollar amount is fixed and the limits are impermeable. Other times, as discussion takes place in general terms about what might be accomplished within the stipulated dollar limits, the amount of the evaluation budget can be negotiated upward (or perhaps in very unusual cases, downward).

The budget document, which should be viewed as a part of the contract, may also vary in complexity depending on evaluation size. As previously noted, in instances of a fixed-price contract, that dollar amount should be stipulated along with the *schedule* indicating when *payments* will be made and under what conditions. By this, we mean that the contract amount might be paid at scheduled intervals (e.g., monthly) or tied to the completion of stipulated activities in the evaluation plan (e.g., reaching agreement on a logic model, conducting the implementation evaluation, collecting evaluation data).

The *basic budget* has as its prime components the costs associated with evaluation personnel, including salaries and employee benefits. Also included is a breakdown of various supplies and materials, equipment, travel, and consultants. A final category of most simple budgets is "indirect costs." This latter item depicts those costs that cannot be specifically stipulated (e.g., the cost of shared office space and various incidentals). An evaluation consultant may have several projects so there is the need to prorate the cost of the space among the different projects. A basic budget is depicted in Figure D.1.

A somewhat more complete and formal budget will provide even greater detail. In this case, the basic budget (shown in Figure D.1) would be supplemented by a budget justification document. This is simply a listing of the various budget items and an explanation of how each was calculated. Thus, personnel are named, their salaries are given, and the percentage of time they will be working is indicated. Also, the basis for determining employee benefits would be described—for example, in some cases, the organization might have a specific health and benefits policy. Furthermore, where travel is an item, policies such as reimbursement rates should be indicated as well as the justification for travel.

More substantial and complex budgets indicate the personnel cost by task (i.e., activity categories of the evaluations). For each evaluation task, staff members are listed with an indication of the time and salary and benefits associated with the performance of the various evaluation tasks. This is depicted in Figure D.2.

January 1, 20____ through December 31, 20____	
I. Direct Costs	$_____
A. Salaries and Wages	$_____
B. Employee Benefits	$_____
C. Supplies and Materials	$_____
1. Office Supplies	$_____
2. Equipment Rental	$_____
3. Telephones	$_____
4. Mailing	$_____
5. Printing	$_____
D. Equipment	$_____
E. Travel	$_____
F. Consultants	$_____
Total Direct Costs	$_____
II. Indirect Costs	$_____
Total Cost	$_____

FIGURE D.1. Sample budget.

RECAP—SECTION D

Contracting for Evaluations

- Acquiring the Evaluation
  - Internal evaluation
  - External evaluation
    - Sole source
    - Request for proposal (RFP)
- Writing the Proposal
- Preparing the Contract or Agreement
- Developing the Budget

Task #/Personnel	Annual Salary	% Time	Task Salary	Benefits	Total Personnel Cost	Task Total
Task 1: Develop logic model						
• Mr. Jones	_____	_____	_____	_____	_____	
• Ms. Gonzalez	_____	_____	_____	_____	_____	
•	_____	_____	_____	_____	_____	
•	_____	_____	_____	_____	_____	_____
Task 2: Reach agreement on questions						
• Mr. Jones	_____	_____	_____	_____	_____	
• Ms. Gonzalez	_____	_____	_____	_____	_____	
•	_____	_____	_____	_____	_____	
•	_____	_____	_____	_____	_____	_____
•	_____	_____	_____	_____	_____	
•	_____	_____	_____	_____	_____	_____
						Grand Total _____

FIGURE D.2. Personnel costs by task.

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

Consider an evaluation of RUPAS. Assuming that it is a 1-year contract and a $50,000 budget, what are some expectations that could be associated with the evaluation? For example, who is responsible for the management of the evaluation study, data collection, and dissemination of study findings? Which of these activities will be accomplished through a partnership between the evaluator and program stakeholders? Which activities will be the evaluator's sole responsibility? And, which responsibilities are the stakeholder's alone? Furthermore, what resources must be in place to ensure the successful execution of the evaluation study? What are some areas in the contract where a compromise can be reached to strike a balance among stakeholders' information needs, availability of resources, and what is feasible on the evaluator's part at this point in time?

Now consider your responses to these questions in the context of a 2-year contract where $75,000 is available for evaluation per year. How might your thinking and approach change?

Further Reading

Bell, J. B. (2015). Contracting for evaluation products and services. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of practical program evaluation* (pp. 769–797). San Francisco: Jossey-Bass.

This chapter discusses issues to consider when securing evaluation services.

Hawkins, P. (2010). Successful evaluation management: Engaging mind and spirit. *Canadian Journal of Program Evaluation, 25*(3), 27–36.

This article offers guidance on how to manage an evaluation contract once secured.

Horn, J. (2001). A checklist for developing and evaluating evaluation budgets. Available at *www.wmich.edu/evalctr/checklists*.

This is another in the series of checklists available through the Western Michigan University site. Horn's checklist provides a guide for developing budgets.

Stufflebeam, D. L. (1999). Contracting for evaluations. Available at *www.wmich.edu/evalctr/evalcontract.pdf*.

This checklist, related to contracting for evaluation, is very complete—perhaps too much so. Use it as a general guideline.

Quick Reads

1. Melanie Hwalek on Evaluation Contracts
   http://tinyurl.com/hbvjbyd

2. Lycia Lima on Improving Impact Evaluation Planning
   http://tinyurl.com/hb22gup

SECTION

E

Who Are the Stakeholders for an Evaluation?

In Section A, I noted that evaluation is "decision oriented" as opposed to research, which is "conclusion oriented." I further indicated in Section C that while there are multiple orientations toward thinking about evaluation, in this book, I focus on a "use orientation"—specifically, "context-sensitive evaluation." I take the viewpoint that evaluations seek to provide information that is helpful in making decisions, increasing understanding, or improving organizational capabilities. Thus, when I do evaluations I need to be generally aware of what use might occur and who will have a hand in making it happen. These considerations are important in determining the "audience" for an evaluation.

STAKEHOLDERS, NOT AUDIENCE

Now let us step back for a moment and consider the term "audience." I have used this term because it is sometimes used in the evaluation literature, but the term is imprecise; "audience" generally refers to individuals who will be receiving a message—a television audience, the audience at a play, an audience with the Pope. Evaluators are not looking to address an audience. Evaluators want to *engage named people* in the activities of an evaluation so that they will feel part of, and committed to, the evaluation. Evaluators want to *hear from* the so-called audience, not simply address them. Evaluators want to *participate with* the

audience. And so, I prefer to use the term "stakeholders"—those who in some way have a stake or an active interest in the program. We have alluded to some of these people in prior sections of this book.

WHO ARE THE STAKEHOLDERS?

Stakeholders are all of those individuals who have an interest in (i.e., are somehow vested in) the program that is to be evaluated. This includes clients of the evaluation, other program staff, program participants, and others in the organization or community.

There are many people who may claim to be "stakeholders." Think for a moment about all the individuals who might in some way have an interest or a stake in a program. Examine, for example, the multiplicity of potential individuals who constitute appropriate stakeholders for a program at an elementary school. There is a program administrator—the evaluation might reflect on how well he or she is managing the program. There are program staff—the evaluation might influence how they conduct their jobs. There are students—Is the program benefiting them? There are parents of students—Is this a program in which parents want to have their children included? There are special constituencies of parents (e.g., special populations)—perhaps underserved groups—Is this program appropriate for their children? There are the principal and the school administrators—Does the program constitute an appropriate use of school district resources? If the program is externally funded, there are representatives of the state or federal government agency or of a private foundation—Should this program be refunded, or alternatively, reshaped? And there is the community at large—Are there elements of the program that reflect badly upon their community?

FOCUS ON PRIMARY STAKEHOLDERS

By now you most certainly have developed an awareness that many stakeholders can lay claim to being an appropriate individual or group for inclusion in the evaluation process. Indeed, the evaluator can become overwhelmed in trying to satisfy and include all potential stakeholders.

Stop! First, let me make something clear: Evaluators do not deal with stakeholder *groups*. Evaluators work with *individuals* who may represent or be part of different stakeholder groups. I think it is best to work with individuals—specific individuals.

What is an evaluator to do? A key question then becomes: How do we deal with so many stakeholders? Several evaluation writers have suggested that evaluators should recognize that while all stakeholders should be heard and should have input, there is a need to pay special attention to certain stakeholders. There are a number of ways in which these priorities can be established.

Which of these stakeholders do we focus on? We can partially answer this by considering these questions: Who makes the decisions? Who influences how the decisions are made? And beyond that, who is impacted or affected by the evaluation? But mostly, *Who wants to use the evaluation information*? These questions, perhaps, provide insight into the potential stakeholders on whom we should focus.

Who Makes or Influences Decisions?

Consider first the issue of *who makes or influences decisions*. Perhaps the most simplistic view is to assume that those who have *commissioned the evaluation* (i.e., those who hired the evaluator) had a purpose in mind and a reason for asking for the evaluation, and thus are likely to want the evaluation to be an input in their decision-making process—to be important in decision making. Perhaps they are the ones who want to know: How is our program doing? In my experience that is sometimes true, but I have seen individuals in various roles who want answers. Those who commissioned, asked for, or contracted for an evaluation certainly may have been included in that group, but they are not the only ones who make or *influence* decisions. Others who are more engaged in the program might be more appropriate as potential stakeholders. They might be the people who have a real interest in the evaluation, because they are "closer to the action" and want to have input in making decisions to improve their program.

Moreover, in many or most situations, decisions are not formally unilateral—that is, they cannot be made solely by one person—but require *agreement among several (or many) individuals*. Perhaps the head of an organization and the director of the program must reach agreement. Perhaps there are others involved who officially participate in making decisions about the program. There are many others who may unofficially have a strong influence on decisions. Think, for example, of vocal and active community members who, while not making decisions, certainly influence them. Think of program participants who might make their voices heard. And there are many others. Each must be considered part of the potential stakeholder group for the evaluation.

Frequently in small local programs that are funded by state or federal contracts there is an evaluation required. The program director (or

someone in the organization who he or she reports to) hires an evaluator. Is the focus of the evaluation to be on decisions that the program director is going to make or on decisions to be made by the funding agency? Can it be both? Is this possible? It is necessary that external evaluation requirements (and their intents) be complied with, but the relative importance of these stakeholders is a function of the organization of the program, its funding, and what decisions might be made and are most important.

Who Cares about Use?

Making or having the potential for influencing a decision are important prerequisites in determining on whom we should focus, but that is not enough. My philosophy of evaluation—the use-oriented context-sensitive approach to evaluation—helps to guide me in how I focus on which stakeholders to pay particular attention to. I start with a strong belief that the role of evaluation is to provide information that will lead to important changes and improvement in the program. I am guided by the conviction that it is essential for evaluators to do everything in their power to see that evaluation findings are considered within the process of decision making; I want evaluations to be conducted in such a manner that they will play an important role in program improvement.

Guided by that belief, in this book I focus more directly on those *individuals* within the system who have a *strong interest* in seeing the evaluation conducted and who might potentially have the power to use, or influence the use of, results in making decisions or in other ways changing the organization. I call these individuals who are likely to be potential users of the evaluation "primary stakeholders." Then I pay *special attention* to the information needs designated by these individuals. Why only primary stakeholders? Quite simply, it is impossible to attend appropriately to too vast a group of stakeholders. I want to focus on those who care and who will likely want to use, and be able to use, evaluation findings as a basis for considering potential program improvements.

How do you find primary stakeholders? As the evaluator, you need to be attentive to seeking out and gaining access to these primary stakeholders—individuals who might use the evaluation to make program improvement. Some who seem to be primary stakeholders may be thrust upon you—you have no choice. For example, the program director or other individual who contracted for the evaluation. Obviously, you must be attentive to his or her needs and directives. But, if that individual's interests appear to be only superficial—he or she has but scant concern about evaluation information and making

decisions—then your attention to that person is somewhat lessened. Likewise, my research on factors associated with evaluation use directs me to assess the *influence capability* of potential stakeholders. This does not mean that we want to include only those who are in a position to shape decisions, but rather we must consider whether particular stakeholders can *influence* potential use.

Frequently, the report commissioners (those who contracted you to do the evaluation) must be viewed as the "gatekeepers." They control access to those on whom you might focus your attention. Try to talk with them about who else might be interested in and potentially use the evaluation information. Indeed, persist in seeking out the names of those potential users. In doing so, you would have implicitly gained their permission to proceed in approaching those who might be—in the truest sense of my definition—real primary stakeholders. These are the people on whom I wish to focus the major part of my attention.

Other Interested Stakeholders

A cautionary note, however: It is important to note that I recognize that these primary stakeholders do not constitute the full group that the evaluator must consider. The professional evaluator still must be sensitive to the variety of *other interested stakeholders*. Talking with, and understanding the perspectives—the culture—of stakeholders who may, for example, lack the power to be actively engaged in structuring the evaluation is a major evaluator responsibility. It is important to ensure that the broader stakeholders' interests are somehow brought to the attention of the primary stakeholders, and their *concerns* are *potentially reflected* in the evaluation.

Why? Because evaluation use is not limited only to major decisions about the program. When program staff members gain a better understanding of the program and that impacts the way they do their job, *that is use*. When those who engage in, participate in, or observe the evaluation process gain appreciation for and understanding of evaluation such that it enhances their potential to improve the program now or in the future, *that is use*. When parents gain confidence in the worth of the program leading to greater support, *that is use*.

DIFFERENCES IN STAKEHOLDER PARTICIPATION

I have an expanded notion of the ways in which primary stakeholders use evaluation. I recognize that when evaluators can more extensively engage specific stakeholders in actively participating in conducting a

study, along with the evaluator, then there are added benefits. The most important is that engaging stakeholders in the process helps to teach them to be evaluation users! In that way, engagement in the evaluation assists in transforming the organization. A further benefit of this expanded participation by stakeholders is an increased knowledge of evaluation and increased appreciation of it. This may lead to future acceptance and receptivity to evaluations, and possibly increased capacity to conduct evaluations themselves.

Now consider the kinds of potential relationships that evaluators might have with various stakeholders. In fact, evaluators interact with stakeholders in many different ways. A thoughtful evaluator wants to assist stakeholders in framing what they consider to be the important evaluation issues, or questions to be examined. Sometimes the thoughtful evaluator wants to be assured that stakeholders understand the evaluation design—the set of procedures that the evaluator will employ—so that they will be reasonably confident in, and accepting of, the accuracy and appropriateness of the evaluation findings. Sometimes the thoughtful evaluator will want stakeholders to participate in instrument development. Stakeholders will usually play a role in data collection and analysis. Many times the evaluator will want stakeholders to assist in examining findings and determining their implications and potential use. Also, the evaluator needs to involve stakeholders in determining how findings are to be reported and how better use of the evaluation information can be attained. (Note that each of these and other activities will be discussed in subsequent sections. See all that you have to look forward to?) And on, and on.

Evaluation does not get completed all at once. There are many *stages* in an evaluation—we can conceive of the involvement of different primary stakeholders in an evaluation in relation to the different stages or steps that an evaluator pursues. Individual primary stakeholders will differ in the extent to which they participate in the various evaluation activities.

Now consider for a moment all of the groups from which primary stakeholders might have been drawn: program directors, organization heads, program staff, program participants, funding agency representatives, and community members. Not every primary stakeholder can be engaged in each stage of the evaluation. Professional evaluators disagree on which stakeholders can be actively engaged in each of the stages of the evaluation. I say "actively engaged" because, of course, it is possible to get general impressions and viewpoints from others prior to each activity.

I recommend that you consider not only which stakeholders to include at each stage of the evaluation but also how deeply these stake-

holders are to be involved in the activity. For example, the evaluation design activity, which is more technical and involves a lot of intense back and forth, would include a more limited group of stakeholders. I refer to this as *deep involvement*. There are many other aspects of the evaluation where it is more appropriate to involve many stakeholders (primary and others) but in a less engaged manner. This is *broad involvement*. Think carefully about what is best in each stage of the evaluation you might conduct.

RECAP—SECTION E

Who Are the Stakeholders for an Evaluation?

- Stakeholders, Not Audiences
- Who Are the Stakeholders?
- Focus on Primary Stakeholders
  - Individuals, not groups
  - Who makes decisions?
  - Who cares about use?
  - Input from other stakeholders
- Differences in Stakeholder Participation

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

Let us begin to consider who the key players are in the RUPAS example. Who do you believe to be the primary stakeholders—that is, those who are in decision-making positions and are most likely to use results of the evaluation? How do they know each other? What is the nature of the relationship that they share? For example, are they longtime friends or do they interact strictly in the capacity of coworkers?

Are there additional stakeholders who would be invested in the program and its evaluation? Who might those individuals include—parent leaders, perhaps? What are their views about the program? How do they experience it? In what ways might their perceptions of and previous experiences with the program influence the evaluation?

Further Reading

Bryson, J. M., Patton, M. Q., & Bowman, R. A. (2011). Working with evaluation stakeholders: A rationale, step-wise approach and toolkit. *Evaluation and Program Planning, 34*(1), 1–12.

This article offers suggestions on how to identify and engage with stakeholders in a manner that would enhance evaluation use.

Mathison, S. (2001). What's it like when the participatory evaluator is a "genuine" stakeholder? *American Journal of Evaluation, 22*(1), 29–35.

You will find an analysis of the roles that evaluators and stakeholders play during an evaluation and the tensions that arise in the process in this reflective piece.

Mathison, S. (2011). Internal evaluation, historically speaking. *New Directions for Evaluation, 132,* 13–23.

This paper provides a historical view of the development of internal evaluation and internal evaluators in the United States.

Taut, S., & Alkin, M. (2010). The role of stakeholders in educational evaluation. In B. McGraw, P. Peterson, & E. Baker (Eds.), *International encyclopedia of education* (pp. 629–635). Oxford, UK: Elsevier.

This chapter provides a particularly good description of the distinction between deep involvement of a few stakeholders and less active involvement of a broad range of stakeholders.

Quick Reads

1. Dani Rae Gorman and Angela Nancy Mendoza on Creating Value and Utility: Engaging Stakeholders throughout the Evaluation Process
   http://tinyurl.com/jca46rp

2. Rena Matthews on Getting Names Right
   http://tinyurl.com/hpp7qob

3. Alberta Mirambeau on Program Stakeholders and Evaluation Stakeholders
   http://tinyurl.com/zkfnb2e

4. Bonnie Richards on Getting Yourself in Context and Developing Stakeholder Buy-In
   http://tinyurl.com/glgz8pv

SECTION

F

How Do You Strengthen Relationships with Stakeholders?

Conducting an evaluation involves far more than just using the technical tools of evaluation. Having this kind of evaluation expertise is important, but the ability to engage stakeholders is, perhaps, just as essential. After all, isn't that ability a necessary part of all that we do? So, let me be specific. Evaluation is a human activity involving social interaction. Evaluators do not work in isolation of stakeholders. Indeed, evaluators and stakeholders must work together to achieve results that will be used.

As we discussed in the previous section, the group of potential stakeholders is very broad. We focus here on "primary stakeholders," which includes those individuals who can influence decisions and who care about use, but as I have noted, there are other stakeholders—program staff, leaders in the larger organization that encompasses a program, program recipients or beneficiaries, community members, and funders. While the nature of the evaluation—the way in which evaluators choose to conduct the evaluation—may influence the extent to which there is contact between evaluators and these various individuals, even the most research-like evaluations involve interpersonal relations with a variety of stakeholders.

There is no simple formula for attaining a positive relationship with stakeholders. Indeed, as I reflect upon my experiences, I note great differences in the extent to which I was able to engage positively in different situations, and considerable variation in the means used to develop those positive relationships. I do, however, note some impor-

tant components that were present and which helped to frame the relationships.

CULTURAL UNDERSTANDING

In my view, the starting point for interpersonal relationships is understanding the culture. Stakeholders are people. You should come to understand and then reflect upon the particular positions and *points of view* of individual stakeholders with whom you are dealing. Where are they coming from? How do they view the program? What are their values, beliefs, practices, and so on? Understanding these—their cultures—will ensure that better communication takes place and that professional collegiality is maintained and enhanced.

These stakeholders work within organizations; this broader culture defines the existing roles, relationships, and responsibilities of those within the organization. Often you may hear, "it is the way that things are done around here"—positive interpersonal relationships are shaped by developing cultural understanding and honoring the prevalent culture. Make sure that you understand the program and its issues.

I will talk in Sections G and H about the program and the organizational and community contexts. Implicit within that discussion is the topic of politics—the need for the evaluator to be sensitive to the political context. But I refer here to more than that—the topic is the development and maintenance of ongoing relationships with people, specifically primary stakeholders. Maintaining a positive relationship with stakeholders does not, in and of itself, make an evaluation successful, nor does failure to do so make an evaluation bad. However, positive stakeholder relationships can have a significant impact in either direction.

First, an important step in attaining this cultural competence is to become aware of one's own culture and how it might influence the assumptions you make about the evaluation situation and the various stakeholders. Who are you? What is your background? What are your values and beliefs? What are your biases—particularly as they may relate to this program? Reflect carefully on this.

CREDIBILITY, RESPECT, AND TRUST

Understanding the culture is the framework—it is both a guiding principle and the first step in gaining credibility, respect, and trust (CRT)—which I consider to be the hallmarks of positive interpersonal relation-

ships. Each of these heightens the possibility of success in an evaluation. The first of these is *credibility*. As I have noted in the research I have done on this topic, there is an initial perception of credibility, but it is further acquired and magnified by our actions. In essence, credibility refers to the extent to which the evaluat*or* is believable, and in turn, the extent to which the evaluat*ion* is believable. As an evaluator, you bring to your assignment particular expertise and experience. This expertise and experience—particularly the perception of it—is what defines initial credibility. And when you, as the evaluator, have relatively limited experience and expertise, the task of acquiring credibility is even greater. Your cultural understanding and the way in which this leads you to engage and conduct yourself can lead to acquiring or enhancing credibility; as you demonstrate expertise, your perceived credibility is enhanced. Some may think that charisma is an important part of credibility, and it is, but in something as long-lived and intense as an evaluation, charisma needs to be backed up by action—positively building relationships.

Respect is both a goal and the means of attaining that goal—that is, showing respect is an important factor in gaining respect itself. The definition of respect provided by the dictionary is "worthy of high regard." To gain respect you must show respect to others. You do this by active listening to what others have to say and by inclusiveness—welcoming the views of others. Remember, you do not have to be liked (it's helpful if you are) but you should strive to be worthy of respect.

The dictionary refers to *trust* as to "place confidence in." You strive to have stakeholders trust the relationship. If an evaluator has credibility and is respected, then indeed, stakeholders will place confidence in his or her actions and judgments. As you can see, there is a great deal of overlap among the goals of CRT. Perhaps I should not have considered these three goals as distinct from one another, but instead asked you to consider them as part of a general pattern—CRT guided by cultural understanding.

SOME PRACTICAL GUIDELINES

I next discuss several additional practical suggestions for attaining CRT.

Professional Image

Maintaining an appropriate professional image is an important element in attaining CRT, which is key in developing better stakeholder relationships. Professional image depends on showing that you are

reliable, trustworthy, and prepared (sounds like the Boy Scout motto—but that's not bad). In essence, you want to show that you do exactly what you said you would do; what you said you would do should be transparent. Being devious and manipulative is not healthy for building relationships.

Aside from being prepared and accountable, another important element is being timely. Immediate feedback or *timeliness* (whether it is in presenting reports, returning calls, or going to meetings) is important. Your credibility is enhanced by your responsiveness, and you gain credibility as someone who cares and wants to be responsive by being timely. Furthermore, unresolved issues should not be allowed to fester—lack of a timely response allows it to do so. I must admit that I'm a bit "over the top" with respect to timeliness: I personally meet all deadlines and arrive early to meetings. I even lower the class participation grade of students in my class for excessive tardiness. My feeling is that when students come late they are disrespectful of the class, to their classmates, and to the instructor. (Aren't you glad that you're not in my class?) Timeliness is important: You don't have to be quite as obsessive as I am.

Cooperation

Being cooperative is certainly a part of professional image, but I think that it is so important that it deserves a separate category. As an evaluator, you should be viewed as someone who is interested in *cooperating* and helping to attain the common purpose of improving a program. Cooperation implies a commitment and a willingness to be part of something. An essential part of cooperation and collaboration is the willingness to accept diverse views. Furthermore, no one is perfect—be tolerant and sympathetic of imperfections in those with whom you interact. And on the other side, accept and acknowledge constructive feedback of your behavior, actions, or views.

In a way, cooperation implies creating a sense of collegiality. You will be working together on a team effort, so to the extent possible, you want to display collegiality and build rapport and not be viewed as a hostile "other." Please note, however, that collegiality is professional in nature. You want to be seen as someone committed to assisting in the improvement of the program—that is your focus. Collegiality does not mean being a "buddy" or a "drinking pal."

Communication

Obviously, communication is an important part of building positive stakeholder relations. How do you interact? You socially interact, you

talk, you listen. Good communication is essential to building trust. When communicating, remember to have respect for others and be understanding of their culture. Each person has unique values, traditions, and experiences. Try to understand and be sensitive to these in your communication. Furthermore, your own communications should convey *clear messages*. Simple language that is understood and appropriate for stakeholders will add to your credibility. Important ideas can be communicated without massive technical verbiage or pedantic language. Hopefully, the style of this book has convinced you of that.

A good communicator is a *good listener*. You will want stakeholders to feel that their views are being heard. You must display a desire to understand what stakeholders have to say in words or what they communicate in their gestures and expressions. Active listening and a demonstration that you have heard and understood a different perspective can be shown by restating a stakeholder's comments in your own words.

Careful observing is also a part of the listening process. Much can be learned and the subsequent relationship enhanced by thoughtful observations. Nonverbal behavior of a stakeholder is important communication, and if understood, can enhance relationships. When you listen and observe, try to keep an open mind. We all filter what we hear through our own preconceptions, experiences, and cultural biases. Try to be aware of how your views are influencing what you hear. The way in which you conduct yourself in personal communications is important. Be civil. (We could use some of that in our legislatures as well.) Remember that you are usually dealing with another professional, and it is appropriate that you show respect for his or her expertise and personal experience. Try to maintain patience, listen to what stakeholders have to say, and whether you agree or disagree, "keep your cool." Above all, don't become hostile or display anger or annoyance.

Finally, let me add a caveat. It's fun to have fun, but *use humor with care*. One person's humor may be construed as another person's insult. Humor may offend people's ethical, religious, or ethnic sensitivities. Humor may be taken personally.

Let me note that I talk more about communication in Section U—"How Are Evaluation Results Reported?" In that section, I talk more about written and oral reporting and the nature and quality of good writing. The main underlying theme of that section (and a lesson for this one as well) is sensitivity and understanding of stakeholders.

Personal Image

How you present yourself most certainly has an impact on potential relationships with stakeholders. Obviously, you can't do much about

your height, weight, or build. Nor can you change your ethnicity or your sex. I feel that attending to many of the issues discussed above can compensate for any possible negative views by stakeholders of your personal image.

There is one personal image aspect that is controllable: You can present yourself as neat and clean. Also, how you dress affects how you are perceived; maintain a manner of dress consistent with the context. Don't dress in a way that says, "Look at me and my fancy (expensive) clothes." Nor should you arrive on the scene in clothes that are too casual or shabby. The advice that I give to my evaluation students is "When in doubt, dress Macy's."

RECAP—SECTION F

A Framework: Positive Stakeholder Relationship

- Development and Maintenance of Positive Ongoing Relationships
- Cultural Understanding
  - Understanding points of view
  - Credibility
  - Respect
  - Trust
- Some Practical Guidelines
  - Professional image
  - Cooperation
  - Communication
  - Personal image

CONCLUDING NOTE

Maintaining positive interpersonal relationships is an important topic. But, as you can see from the above discussion, it is a difficult one to grapple with in terms of any kind of precise description of what is required. For that reason, the topic is barely dealt with, if at all, in most evaluation textbooks. I believe that many important components have been touched upon in this section and I urge you to consider these things when you conduct an evaluation.

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

Let us continue with our evaluation of the RUPAS program. Understanding stakeholders' perspectives and prior experiences with evaluation will help in contextualizing and guiding your interactions with them. Do we know how stakeholders—such as Amy Wilson, Mary, Zoe, Carmen, and Children's Trust—have previously experienced evaluation? That is, have they actively sought out evaluations? Or have evaluations historically been done to them? In the same vein, since no two persons' opinions are the same, what might each of these stakeholders view as the evaluation's purpose? What questions can we ask each individual or group so that we can learn more?

Building relationships is as much about getting to know your stakeholders as it is about allowing them to get to know you. In these early discussions, also think about those aspects of your résumé that might help you to connect with them. Do you have any personal experiences that enable you to better relate to the program context and to the various stakeholders? What might they include? Do you, for example, have previous nonprofit experience? Or do you perhaps have family that are migrant workers?

Further Reading

Adams, A., Nnawulezi, N., & Vandenberg, L. (2015). "Expectations to change" (E2C): A participatory method for facilitating stakeholder engagement with evaluation findings. *American Journal of Evaluation, 36*(2), 243–255.

This article describes an approach for supporting stakeholders' use of evaluation findings.

Bryson, J. M., & Patton, M. Q. (2015). Analyzing and engaging stakeholders. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of practical program evaluation* (pp. 36–61). San Francisco: Jossey-Bass.

The emphasis in this chapter is that stakeholder analysis is an essential first step in attaining stakeholder engagement.

Fierro, R. S. (2016). Enhancing facilitation skills: Dancing with dynamic tensions. *New Directions for Evaluation, 149,* 31–42.

Common challenges that arise in the course of evaluator–stakeholder interactions are discussed here. Suggestions for improving these interactions are offered.

Stevahn, L., & King, J. A. (2016). Facilitating interactive evaluation practice: Engaging stakeholders constructively. *New Directions for Evaluation, 149,* 67–80.

This article defines interactive evaluation practice as a basis for decision making and conducting a meaningful evaluation.

Quick Reads

1. Caroline DeWitt on Valuing Stakeholders in the Evaluation Process
   http://tinyurl.com/hvpflce
2. Clara Hagens, Kelly Scott, and Guy Sharrock on Embracing an Organizational Approach to ECB
   http://tinyurl.com/zt2uop7
3. Chari Smith on Assess Their Attitude toward Evaluation before You Get Started
   http://tinyurl.com/hkysg7e

SECTION

G

How Do You Describe the Program?

One of the first steps in performing an evaluation is gaining a clear description of what constitutes the program. But first, let me talk about what I mean by "program." The dictionary definition: A program is a complex of people, materials, and organizational structures having a particular goal and serving a particular population. This is a reasonable starting point—we discuss it more fully and then explain how you might go about finding information to learn more about the specific program you will be evaluating.

WHAT IS A PROGRAM?

Let's start by considering a program as having a purpose or *goal,* because that is, in fact, the most important aspect of a program's description. Indeed, the noted management consultant Peter Drucker points this out nicely in the way that he approaches work with major corporations. Can you imagine Professor Drucker in his initial meeting with the management of General Motors and asking them, "Would you please tell me what business are you in?" You might imagine the astonished look on their faces. But what he is asking for is a precise definition of the organization's goals.

In practical terms, we want to ask, "What does this program hope to accomplish?" "Why does it exist?" Clearly, a third-grade mathemat-

ics program, for example, exists to teach children mathematics—the mathematics that is generally acknowledged to be at a "third-grade level." However, there are differences in what are believed to be appropriate outcomes for teaching mathematics in the third grade. Perhaps the goals of the program are tied to achievement on standardized tests. These goals help shape the program. Perhaps also, success on standardized tests is only a part of the goals of the program and other outcomes are also considered to be important. What other goals are there? Perhaps the community and the school district have other considerations that they deem important—facilitating everyday understandings of mathematics, or liking and feeling comfortable with math.

It is relatively easy to talk in general terms about why a program exists—what it hopes to accomplish—but as the old saying goes, "The devil is in the details." What specific behaviors are to be modified? What specific attitudes are to be changed? If there are multiple objectives (multiple important desired outcomes), it is important to know whether they are of equal worth. Do certain areas play a greater role in informing an overall judgment about the program? The evaluator might gain some insights into the goals and objectives deemed to be important by examining performance measures currently in use (if there are any). One cannot assume, however, that the measures in place necessarily match the intended objectives for the program. There may, in fact, be many minor and unstipulated goals. All of this is to be determined.

Another element in understanding the program involves the determination of *who is to be served* by the program. In the third-grade mathematics example, the initial answer might be "third graders." However, that is simplistic. Are there unique characteristics of the third graders that should be understood in developing an evaluation? Which third graders are to be served? Do we want to know about how different subgroups of students are performing? How were students selected to participate in this program? Were some third graders at this school included, or were all of them included? If some third graders were selected, which ones participated and how did they differ from the full complement of third graders? Alternatively, perhaps some students were selected from a variety of third-grade classes to participate in an after-school activity. Again, it is important to know how students were selected and how they differed in systematic ways from nonparticipants. Identifying who is served by the program and accurately describing them is an essential part of understanding the program. It is not as easy as it might seem.

Programs are further defined by the *general approach* they take in working with clients (those served) to achieve the desired goals. This

general approach consists of services and structured activities. An important part of understanding the program is defining the unique set of activities and procedures that the program purports to provide. What is the schedule of activities? Do activities vary from day to day? Is there a plan that defines the activities specifically?

For example, my project team conducted an evaluation of an after-school arts project that teaches mathematics concepts. In this program, third-grade students received additional classroom instruction in math and literacy using an art-based curriculum. Specifically, they participated in 1 hour of journal writing activities connected to art for 12 weeks, followed by 1 hour of math activities related to basket weaving for another 12 weeks. During the math component of the program, students were taught measurement skills and used geometric concepts to create their baskets.

Programs are run by people. There are *individuals* who do the work. Perhaps they instruct, perhaps they counsel, perhaps they provide health or rehabilitative services, perhaps they provide transportation. Usually, there are people who administer; who ensure that the program activities are occurring. We need to consider who these personnel are and what role they play in producing the program services that the organization provides. We need to understand better who these people are in terms of the special skills, aptitudes, or training requirements that need to be present in order to perform their required function. It is also helpful to understand something about their values and their individual beliefs. Individuals differ in their belief system and values—their culture. This applies to staff members and to all stakeholders. (We discussed this to some extent in Sections E and F.)

Programs often employ *physical materials* needed for activating the program. This might include textbooks, workbooks, computers or computer software, or various specialized materials. These should be identified since they help to describe the program.

Programs may also differ in *size*. Think for a moment about the great variety of entities that might be considered a program. In an elementary school, there might be a special program to enhance mathematics learning by engaging children in doing art projects that involve mathematical concepts. Or the whole third-grade mathematics teaching activity at a school might be considered a program. One might also evaluate the entire third-grade educational program of a school or school district, including the various subject areas and multiple outcomes. Or a state government might conduct an evaluation of third-grade mathematics statewide. Programs come in different sizes and are typically components of larger organizations, which themselves might be considered programs.

Programs that are to be evaluated also have different *developmental stages*. It seems that we typically think of "new" programs as those that need to have their worth determined. For example, picture an alcohol abuse program that has been operating for some time and people are generally happy with it—or perhaps not happy with it. A new concept appears in the literature that seems to promise a more effective way of achieving the same goal. A new program is then established—to be tested and to be evaluated. What is asked: Does this approach that we read about in the literature, and are now implementing on a pilot basis, really work? Or more precisely, does it really work better for us than what we are doing now—*for us, in our situation, in our location*?

However, programs to be evaluated need not be new. One could simply say, "We've been doing this program for some time. It is our regular activity. We believe that we are satisfied, but perhaps we ought to verify that it is working as well as we think it is." We simply would like to know, "How are we doing?" Unfortunately, that occurs all too infrequently. We more often accept what exists and seek only to evaluate what is "new."

Whether a new or a continuing program, it is important to be aware of the developmental stages of the program. The extent of the development and implementation of a program will set the parameters for what might be accomplished in an evaluation. For a newly established or early-stage development program it may be inappropriate to consider the outcome measures. The program is just getting its feet on the ground and you may need to focus on getting it to run (or walk) better.

Programs to be evaluated might not differ in their activities but might involve intended *changes in organizational structure*. We tend to think of programs as service providers (e.g., teaching math, preventing substance abuse), but programs might also deal with modifying organizational structures or functions. For example, within a county social service agency, a program might involve a reshuffling of responsibilities to different constituent agencies—that is, in an attempt to challenge the traditionally segmented provision of social services, a social service agency might centralize all services under one roof. Or a program might involve a change in reporting within the organization that administrators believe will increase efficiency. Or staff responsibilities might be modified with different staff members taking on added responsibilities (or different responsibilities altogether).

I noted earlier that it is important to also clarify what is *not in the program*. The program to be evaluated could be part of a larger entity that has multiple programs. In that instance, it is important to identify which aspects are uniquely part of the program being evaluated

and not of adjoining or encompassing programs. It is also important to identify which personnel are part of the program being evaluated. For staff who are part of multiple programs within the same organization, it is necessary to identify the portion of their time and the nature of activities that are not part of the program to be evaluated. The same goes for materials, equipment, procedures, and all other aspects of that which constitutes a program.

Sometimes a program activity is part of a larger program, or it could be a "designed supplement" to an ongoing program. The program to be evaluated might occur simultaneous to the larger program or it might, for example, be a follow-on entity. In that case, it is essential to understand which features are part of the regular program and which are uniquely a part of the "add-on."

LEARNING ABOUT THE PROGRAM

Now that we have talked about "what is a program," let us consider how you, as the evaluator, can begin the quest to understand the program to be evaluated. Remember, no matter how technically skilled (or not skilled) you are as an evaluator, *you must first be a learner*. You should try to learn about the program by both examining written documents and interviewing individuals who can help to provide insights. These are not necessarily sequential activities, but do, in fact, involve continuous interplay. You might learn from documents; you could get clarification in doing interviews; you might consult documents again to further elaborate your understandings; and, finally, perhaps talk some more. Let's consider each of these in turn, remembering that there is a back and forth between them.

Documents to Review

The first step in learning about the program is examining the various program documents. At this point, you will not be seeking to understand the success of the program, but rather, the *intent* of the program—what the program developers say it is and what they say it does. There are five main types of documents that might be examined: (1) the written program proposal and materials related to the proposal, (2) guidelines of the funding agency, (3) program materials, (4) management documents, and (5) past evaluation reports (if any).

Most important of these is the *program proposal*. In many instances, a written proposal has been prepared to either obtain funding for the program or to authorize its establishment using existing resources.

A proposal certainly would be available when the program has been externally funded. In instances where the program has received funds from an external source, proposals might have been submitted to a state government, a federal government, or a charitable foundation, among others. In those cases, the program designers were required to specify goals of the program and what it hoped to achieve, the resources that would be employed, and the manner in which the program would operate. If the program was established within the organization, and with new internal funding or a reallocation of funding, there most likely would have been a plan or description of what the program was intended to be. In this instance, staff and other individuals within the organization considered the need for doing something new—modifying their existing procedure—and wrote documents indicating what it was that they hoped to accomplish with this newly reformed program and how they intended to do it.

What information can you obtain from the program proposal? Clearly you will want to know what were presumed to be the *goals of the program* (e.g., reducing obesity) and the specific *objectives* that are presumed to lead to that goal (e.g., learning about different food values and calories).

You will also want to know the names given to particular activities that will be engaged in for acquiring such skills (a particular curriculum). As you dig deeper, it is necessary to obtain further specific understanding about the *program activities*. Simply knowing the name of an activity is not enough. Names have many meanings, so establishing the particular meaning attached to the activity's name is essential. Next, you need to know when exactly each activity will occur, for how long, and with what time line: Are there specific program materials that will be used (e.g., computer programs, handouts, status assessment sheets)? Are there particular engagement strategies that program staff will employ (e.g., lectures, peer interaction, counseling, empathy, companionship)?

Furthermore, you will want to discern the *intent* of those who developed the proposal. Developers might have had various working meetings or enlisted work groups to consider various issues. If documents from such meetings are available, they would be a helpful resource. You want to know what people were thinking when they developed the proposal. This is comparable to trying to know the legislative intent behind particular bills passed by Congress.

It is important to strike a certain cautionary note here. Goals, activities, and intents as specified in the program proposal do not always ring true. Primary stakeholders may not have fully agreed with the program proposal. Moreover, in the natural course of events the pro-

gram may have undergone change. Looking at the program proposal provides a historical context. Subsequent examination of materials and interviews will refocus your program understandings.

Another source of important documents in understanding a program are the written *program announcement materials* that may have been sent to various constituencies describing the new program. A new program having been established, the program developers wanted to "announce" its arrival. Staff had to be informed that the program had been approved along with being provided a reminder about the program characteristics. The documentation might include e-mails, text messages, or flyers. There would also have been a need to inform clients (or potential clients) of the program. Thus, brochures or other materials describing or advertising the program might have been prepared. Finally, the community at large could have been informed of the program through a variety of the above information sources, but might also have been communicated to by a newspaper article. All of these are avenues for learning about the program.

The various *program operational materials* are documents to be examined in seeking to understand the program. You can obtain enhanced understanding of how the program operates through examining the various materials that are used as a part of the program. The program may have handouts that were distributed to clients, including instruction sheets or worksheets. In some cases, these may simply provide descriptions of intended activities or expectations. In other instances, these materials are the documents that are used as part of the particular activities of the program. Also, it is important to examine the forms used to record program participation.

Program materials include various *management documents*. For example, you might gain greater understanding of the program by examining a list of the staff and their program responsibilities. The management documents might also provide information on the minimum competencies required of staff. Is there a specific language competency necessary for staff? Is there a need for specialized training related to the nature of the project? Is there a particular academic background that is required? What kind of cultural competencies are needed? What kind of prior experiences are newly appointed staff expected to have?

Along with the staff list, it is helpful to obtain an *organizational chart,* if one is available. Who is in charge of the program? What are the formal staff-reporting responsibilities? Do some staff have direct reporting responsibilities? Are there staff who report to someone outside the organization? (But note: Formal reporting charts do not always depict what really transpires.)

Finally, an important program management document is the *budget*. This provides insight into the costs associated with various personnel and the costs of various program materials. Also essential is an understanding of the various other program expenses; the amount allocated for evaluation is an important item. You bet!

The fourth major kind of program material is *prior evaluation reports*. If the program had previously been evaluated, then obtaining past evaluation reports will provide insights about the program that will aid in subsequently fashioning an evaluation. What you primarily want to know from previous evaluation reports are comments about program implementation. Had the program been implemented in the way in which it was described in the program plan? What might the reports tell you about the operation of the program? Are you in fact looking at a program that differs in substantial (or minor) ways from what it purports to be? Examining the evaluation reports helps you to understand what the program was when it was last evaluated. It may be different today, but it is helpful to obtain the written picture provided by prior evaluation reports. Viewing past achievements is of less concern because you will be focused on the current accomplishments of the program. However, viewing past achievements may well offer clues to areas that require the evaluator's attention.

Interviewing Stakeholders

Aside from documents, interviews are a key part of the process of getting to understand the program. A substantial amount of informal interviewing takes place at the very early stages of reaching agreement about the evaluator's hiring or participation in the evaluation. You might think of this as the "getting-to-know-you" stage. This, in itself, is very important because the building of a relationship with critical individuals in the program is the first step in building evaluator credibility, but it is also the start of getting to know the program. It is important to identify the individuals who are most likely to be able to respond to important questions that you might have about the program. In trying to gain understanding, you will need to talk with those who are most knowledgeable about the program, whoever they may be. Generally, this would include a meeting with the project director. Also, key staff might be the foci of your interviews. Administrative personnel within the larger organizational context can often shed light on the program and its goals and intentions vis-à-vis the larger organization. Frequently, it is appropriate to include people who are not currently part of the program, but who had previously been instrumental in developing it. This would include the grant writer if this was an externally funded evaluation.

Now to continue, I would *not* include clients or community at this stage of the evaluation when your concern is gaining program clarification. You, as the evaluator, want to know what the program is supposed to be—that is, how the program *is supposed to operate,* rather than how the program *is actually functioning*, or how some would *like it to operate*. These involvements occur at a later stage of conducting the evaluation.

Much might have been learned from initial interviews with staff and from subsequently examining program documentation as I have previously described. However, many questions will still remain after the documents are thoroughly examined. Additional interviews may be necessary and appropriate in order to clarify understandings about the program. Documentation and interviewing are complementary activities; the evaluator may go back and forth in seeking clarification.

One cannot evaluate a program without fully understanding what the "it" is. Getting clarity on "it"—the program—is an essential part of the evaluation process.

RECAP—SECTION G

Describing the Program

- What We Need to Know about Programs
  - What are the goals?
  - Who is served?
  - What is the general approach?
  - What people and materials are required?
  - What is the program's organizational context?
    - Program size
    - Organizational location
    - Program's developmental stage
    - Programs modifying an organizational structure
    - What is not in the program
- Learning about the Program
  - Documents
    - Written program proposal
      - Goals
      - Activities
      - Developer's intent

- Guidelines of funding agency
- Various program materials
- Management documents
- Prior evaluation reports

- Stakeholder interviews
  - Primary stakeholders
  - Beginning to understand their individual cultural context

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

We discussed the general RUPAS landscape in previous sections by considering the stakeholder context, by trying to understand who they are, and by learning how to build credibility and trust. Let us now begin to think more carefully about the RUPAS program itself. Pause for a second here and ask yourself: What do I know about the RUPAS program? What does it intend to accomplish? Who is to be served? What services are provided? What people and materials are required? What is *not* in the program? Is there anything else I need to know? Who can I ask? Got it? Now let's take a more critical look.

Given the information that you have before you, consider whose views of the program are being represented here—that is, have you heard primarily from program staff? How much have program participants contributed to your understanding of the program? Do you trust everything that you have read, heard, and seen of the program? Are there any gaps in understanding? Is there anything that does not seem to match up? If so, how can you learn more? For example, what questions did you previously ask about the program? Could you ask them in a different manner? To whom did you ask those questions? Are there others with whom you can speak?

Further Reading

Altschuld, J. W., & Watkins, R. (2014). A primer on needs assessment: More than 40 years of research and practice. *New Directions for Evaluation, 144,* 5–18.

This article provides an overview of what a needs assessment is, its role in the evaluation process, and how it can contribute to your early understandings of a program to be evaluated.

Conner, R. F., Fitzpatrick, J. L., & Rog, D. J. (2012). A first step forward: Context assessment. *New Directions for Evaluation, 135,* 89–105.

This paper, as well, raises issues to consider as you are trying to gain clarity on the program's context. It could help you to frame questions to ask program administrators about context.

Quick Reads

1. Jennifer Greene on Context of Evaluation, and the Evaluator as Part of the Context
   http://tinyurl.com/jo5f5kf
2. Monique Liston, Leah Peoples, and Ibukun Owoputi on Evaluating Equity, Diversity, and Inclusion in Organizations
   http://tinyurl.com/h4xksr7
3. Kristin Mendoza on Organizational Culture and Practice
   http://tinyurl.com/gmv572k
4. Michael Quinn Patton on Evaluation and Politics
   http://tinyurl.com/gsj978g

SECTION

H

What Is the Organizational, Community, and Political Context of the Program?

Another part of "understanding the program" is understanding its context. This means understanding as much as you can about the program to be evaluated and all its influences. You might be thinking, "Well, Marv, that's very broad," and you're right. There is a broad milieu—a broad context—that makes up programs and what surrounds them. Programs do not exist in a vacuum and they are not containerized. So how do we know what is part of the program context and how do we go about this process of discovery and learning about the program?

Let's consider this: Programs have an identity. Programs are part of larger organizations. Organizations are part of communities. All of these are made up of people who have unique values—unique cultures. Some people share similar values while others do not, and when values, people, organizations, and communities interact—which they inevitably do—politics and political processes are in play. Thus, understanding the program's total cultural context means learning about the program as an organization, the people in the program, and what is important to them. It also means learning about the community that surrounds the program, the people in the community, and what they value. This sounds straightforward enough and maybe even a bit trite, but context is to evaluation what location is to real estate. It is *that* important. Successfully navigating these various layers of context

is perhaps one of the most challenging tasks that an evaluator must be able to do for the evaluation to be useful (and hopefully, used—more on this in Section V). So let's think about what we call "context" a bit more and what you can do to make your way through this terrain.

ORGANIZATIONAL CONTEXT

Programs exist within an *organizational* context. Part of this structure is the organizational location. What is the governing agency or larger organization of which it is a part? Is the governing agency also the program's sole funding source? In such cases, the governing agency and the program are nearly synonymous. Other times, the governing agency may have a portfolio of programs that are designed to contribute to its social mission. Such programs may be offered by different levels of an agency. They may be implemented in many different geographic locations and may even be funded by different outside entities. All of this influences the program's and the organization's culture—the beliefs, customs, and rules that shape the way the organization operates.

Furthermore, programs, as organizations, have a *history*. This history reflects many previous decisions and involves those currently or previously within the program. It also involves individuals within the larger organization encompassing the program and the encouragement or constraints that they have provided and continue to provide. Consider, for example, an after-school tutoring program that is situated in a school and that school is part of a school district. The program did not simply come into existence at the school. Rather, someone or some group of people designed it to address a specific set of concerns. These individual(s) then had to convince a broad audience—teachers, school leaders, district leaders, and families—that the program was worth considering and supporting. Once the resources had been identified and allocated for the program, it had to be implemented. There is a story behind the development and existence of every program. A part of understanding the program's organizational context is uncovering this story and learning its history.

➢ What You Can Do

1. Do your homework. Learn when the program was established, by whom, where, and why. How large is the program? Who is on staff? Who does the program aim to serve? What are its goals? Much of this information should be available on the program's website or through social media. If the

program is government funded, then you might locate the call for program proposals in the government registries. This document will provide a good sense of the program's focus.

2. Talk with program staff. As you meet with stakeholders to learn about their evaluation needs, take the opportunity to learn about the program as well. Engage staff in an informational interview. You might ask, "Who was instrumental in starting the program? Whose idea was it? What were the motivations for establishing the program? What needs or issues was it expected to address? Has the program been implemented continuously?" If there was a hiatus in the program's availability, why and for how long? Likewise, if the program is to be *newly established,* many of the same questions may also be asked. "Who advocated for the creation of the new program? Who preferred the old program and maintained that a new program need not be established?"

COMMUNITY CONTEXT

Now let's consider the community context. Just as every program has a history, so does every community. Understanding a community's history provides a unique perspective as to how it took shape and, more importantly, why it exists in the form that you will come to encounter and know through your evaluative work. Thus, you will want to understand those aspects of a community that give it a distinct identity and "feel." What do I mean by this?

Let's consider the after-school tutoring program mentioned earlier. How do we come to understand a program that is situated in a rural environment, where gas stations and markets are over 40 miles apart, where children are bused to school, and where bartering might be a common practice? Such a program would be very different from one that is located in a suburban town where the majority of homes look similar, where all lawns appear uniformly green and freshly manicured, where town centers are typical places for neighbors to run into each other while out on errands, and where children can walk or ride their bicycles to and from school. Life in metropolitan and more urban areas provide yet another perspective. What do we make of a tutoring program that is offered in communities where children are strongly discouraged from playing outside after dark; where gunshots, sirens, and helicopter blades beating overhead are reliably heard several times per week; where lockdowns are part of a typical day at school; and where there are more liquor stores than grocery stores per square mile?

It wouldn't be unusual for the same program to be offered in these highly different settings. However, we would have to anticipate that

the manner in which it is implemented and those who are truly able to take advantage of the program would change from one environment to the next. As you can see, every community has a story, and geography is only one element of it. Getting to know the community context requires much more.

➢ *What You Can Do*

1. Do your homework and find out what the story is. Continuing with the after-school tutoring program as an example, you might consider questions such as: What is the economic health of the community? What occupations are primarily reflected in the community? What kind of housing exists? What is the education level of community members? How many languages are spoken and what are they? What is the crime rate over the past few years? Much of this information can be acquired from area demographic statistics, but that is only a start.

Remember, you must obtain a "feel" for the community. One easy way to start gaining a sense of what a certain community is like is by using Google Maps. The "street view" feature of this Web-based resource allows you to walk down a street in most neighborhoods. Thus, without leaving your office, you can see where parks, libraries, or liquor stores are located within different communities. This is only the tip of the iceberg and does not preclude your physically going into the community.

2. Talk with community members. Every neighborhood and every community has people within it. My message for evaluators: Go into the community and get a feel for who these people are. What does a typical day in their life look like? What are their social moods? Do they know about the program? Do they have strong feelings about it? Do they have any knowledge about it?

Also, get a feel for the social context. Have there been changes in the economic structure of the community, such as major companies moving in or out? Has there been civil unrest or increases in the crime rate? What kind of relationship exists between the community and its leaders? Are there some issues that are "electric," or too hot to touch? Why?

POLITICAL CONTEXT

We have talked about the program and its larger organization as well as the community that encompasses it. Now we look at both from a political perspective. Esteemed evaluation writers have said that all evaluations are political. Within evaluation, there exists politics. First, let me

make this clear: Politics, as such, is *not all negative*. Programs themselves are created through a political process. There are political mechanisms that helped create the program and which continue to foster it. At the start, some people had a view about a way to achieve particular goals. Others needed to be convinced of the propriety of the proposed set of actions. Some might have disagreed. There was a give and take and finally, a political consensus was attained. Programs reflect a political consensus—a compromise—and an accommodation of multiple views.

Evaluation has continuing political impact because its purpose and the process are political. Consider, for a moment, what evaluations do. In a formative evaluation, we might question whether the logic behind program activities is sound—whether these particular program activities are capable of attaining the desired goals. In a summative evaluation, the viability of a program's goals is examined. Generally speaking, evaluations are relied upon to drive decision making—whether aspects of the program should stay the same, be altered, or discontinued (more on this in Section V).

A further indication of the political impact of the evaluation process is found in what we evaluators do in an evaluation. We assist in the determination of which stakeholders will participate and in what way, and we jointly come to a decision about who are the primary stakeholders. We obtain input from the larger stakeholder audience. At the same time, we are constrained by how attentive we can be to each stakeholder. Decisions about who gets selected and participates are sometimes logistical, but they are also often political—such decisions are measures of power. Furthermore, since programs were conceived politically, the results of evaluations might potentially disrupt this politically derived agreement about a program. Evaluations could disrupt the power balance that created the program. There are many aspects to this politically accommodated balance—the program being evaluated could have been created within the organization sponsoring the evaluation. However, there may have also been strong community views expressed during the program's creation. Evaluators cannot respond to all stakeholder points of view and represent them equally. Thus, evaluators risk politically antagonizing some groups.

One source of this continuing tension is the different value systems. Value systems influence ways of operating and views on acceptable behavior. They are made up of views and opinions rooted in previous experiences, beliefs, and cultural practices, and are influenced by the social norms of the organizations and other entities in which individuals take part. Every stakeholder group involved in an evaluation—directly and otherwise—has a value system embraced by their own family, friends, and social group. Many evaluators tend to refer to these

value systems as "cultures." There is a community culture, an organizational culture, a program culture, and many individual cultures. These various cultures impact the way that everyone experiences the program and its evaluation along with what is considered acceptable. These disparate cultures are often the root of conflict and misunderstanding because they do not always align. In particular, they may not align with your values—your culture. (Yes, the evaluator has a value system as well.)

Thus, evaluators must be mindful of existing cultures, how they are represented, and how they are experienced. This sheds light on the dynamics that exist between individuals and groups within the evaluation context. Note that the views and opinions that you encounter may belong to either a dominant or a minority subgroup. What do I mean by this? Communities are made up of smaller groups of constituents. Each group has their own views about and experiences with the program. Thus, some views are inherently overrepresented while others are inadequately represented. How will you—the evaluator—represent these competing perspectives? Likewise, while there are commonly accepted value systems, there are also sets of beliefs among segments of a community that are less widespread. This is particularly true for traditionally underrepresented groups whose voices may not get heard because they are intentionally not invited to contribute to the process, or because access, language, or cultural barriers prevent them from fully participating. Again, what role will you play in ensuring that the needs and voices of these groups are accurately represented?

It is absolutely important for evaluators to understand the political context because embedded within it are value systems, power structures, and implicit and explicit expectations (more on this in a bit). This aspect of context warrants mention because it sets the tone for the program, the community, the people in these settings, and how the evaluator comes to understand it all. So, what can you do? The theme of this section is "There's a story. Find out what it is."

➢ *What You Can Do*

1. Do your homework. Consider the community influentials who might have thoughts about the program, its evaluation, or its outcomes. In the case of the tutoring program mentioned earlier, it could be the parents, the school principal, the district superintendent, a city council member, or even the mayor. How would you know who the possible influentials are? To start, pay attention to the social issues that are making local, regional, and national headlines. Do they affect the program in any way? If so, how? Can this be verified with program stakeholders? Find out who is weighing in about the

program, the community, and what do they have to say. In the same vein, explore relevant archived materials. Examples of such resources include proceedings or recordings of town hall or city council meetings. Similar documents exist for school board meetings. They are often publicly accessible on city or school district websites. You might even ask program administrators if they would allow you to peruse prior meeting minutes for your own education.

2. Talk to people. Many in the community may not be aware of the program and so are apathetic about it. But there are voices (both proponents and opponents) who might have views about the evaluation being conducted and these should be considered a political factor. Program opponents, while possibly wanting an evaluation, will certainly have views about what they would consider to be desired results. Strong community advocates of the program might be hesitant about an evaluation because they don't want to see the program changed. Alternatively, they might want an evaluation in order to validate their position.

Various individuals might be impacted by the program in one way or another. For some, the *continuance* of the program might be viewed as beneficial. For others, the *dis*continuance of the program might be applauded. Particular aspects of the program might be viewed as intrusive or controversial. People in the community might think (or say), "Do I want these people [program participants] in my neighborhood? How will this impact the traffic on the streets that I drive? Do I want my son, daughter, partner, or friend participating in such a program? Does this program offend my ethical sensitivities? Are vulnerable constituencies having their voice heard? Are there ethnic or religious issues that I object to? Are social justice considerations being appropriately addressed?" Clearly, not all community views can be taken into consideration. You, as the evaluator, should at least attempt to gain some understanding of these sentiments.

Ask as broadly as possible about those in the community with whom you ought to talk in order to get a better understanding of the diversity of views about the program. And thus, be armed to perceive the sensitivity of what you do and what you report.

IMPACT ON THE EVALUATION

We have discussed the program, community, and political contexts because they are all important aspects of the evaluation context—that is, the setting and climate in which the evaluation occurs. Another important element in this organizational context are the various stakeholders. Some of these may be included in this discussion (a fuller discussion of stakeholder groups is presented in Section E). It is critical to

understand the web of relationships that people have with one another and with the program. The intermingling of unpredictable views and variable opinions within such webs make the evaluation context highly complex. Your ability to conduct a successful evaluation—one in which learning takes place and results are used—is contingent on your grasp of what is happening in this broader environment.

Consider, for instance, how you would go about conducting an evaluation where staff are excited to understand whether participants are benefiting in the manner that is hoped, where staff are open to feedback about the program's strengths and weaknesses, where the program itself is valued and supported by the community, and where there is unilateral commitment to see it thrive. Compare that with an evaluation of a program that is openly disdained by the community, but vehemently supported by program developers and funders, and where evaluation is not an instrument for learning and improvement, but is instead a mechanism for improving public relations and marketing. Likewise, what if you find yourself in a situation where staff are ambivalent about the program and are comfortable with the way things are, where participants' feedback are taken with a grain of salt, and where administrators are more interested in doing evaluation for the sake of being able to say—to a funder or an accrediting agency, for example—that they did it. Some of these examples are admittedly extreme, some are perhaps rare, but they have occurred and every so often, an evaluator will have the opportunity to work in such contexts. The spirit of evaluation differs greatly in all of these cases, and upon deciding that you will pursue an evaluation, it is important that you familiarize yourself with circumstances that may affect your work and trajectory.

➢ *What You Can Do*

1. Again, do your homework. Understand whether evaluation has been conducted in the past, what was its purpose, and who were the intended audiences. Knowing what role evaluation played (if any) with respect to the program's development and evolution will give you a sense of how to go about the evaluation that you must conduct. Sources that might prove to be helpful in ascertaining this kind of information include existing evaluation reports, newspapers, other types of media, government registries, governing board records, and archived materials. Note that evaluation results are often reported in the media, but they are not always referred to as such, so be keen when searching for and consuming information from these sources.

2. Talk to people—program administrators, program staff, program participants, and community members—who you think will be able to shed

light on what can be expected in the present or upcoming evaluation. Specifically, try to understand their previous experience with evaluation, how they have used those results, their motivation for engaging in or commissioning an evaluation, and how they intend to use the findings. Know that there may be program stakeholders in the organization who desire an evaluation and others who would be threatened by one. You need to consider what *motivated* the evaluation. Who asked for it? Who wants it? Does the request come from the larger organization? Do they have an agenda? Are they open to real evaluation? Is there political pressure within the organization to satisfy the demands of a funding agency?

Talking to primary stakeholders to personally gain understanding of context is important but it has value beyond that. Involving these stakeholders as you seek to gain this knowledge of the context deepens their own understanding and helps to attain a more shared perspective.

➢ ***My Advice:*** It is important to know that tension will resurface throughout the evaluation because every evaluation activity has political consequences. Your role, however, is not to force alignment. Rather, it is to display sensitivity toward these dynamics, acknowledge where there may be discordance, critically reflect on whether they affect your ability to conduct a balanced evaluation, and attend to them throughout the evaluation.

Finally, recognize the political reality of evaluation. It exists and encompasses various partisan views, but do not be deterred. Do the best that you can to personally conduct the evaluation in an unbiased fashion. Always be sensitive to the political context surrounding your endeavor.

➢ ***Thinking Ahead:*** In Sections O, P, and Q, I discuss development of the evaluation plan. I urge you to be aware of potential political and organizational issues that might need to be addressed in the plan. By this I mean, begin to consider the manner in which sensitive issues might be dealt with. Consider whether the views of stakeholders in the community or the organization might interfere with the conduct of the evaluation. For example, will access be limited, or hindered, by those antagonistic to the program's continuance?

Being "in the know" about community, political, and organizational issues and points of view adds to your ability as an evaluator to relate to the various individuals who are connected to the program—including stakeholders, program staff, and those whom the program services. This helps you to make sense of what you perceive and of the information you gather. Contextual knowledge will assist in understanding the nuanced occurrences unique to this particular individual program.

RECAP—SECTION H

Organizational, Social, and Political Context

- Organizational Context
  - Research the program's history
  - Talk to program staff and program participants
- Community Context
  - Learn about groups and subgroups in the community
  - Go into the community and get a "feel" for what life is like there
- Political Context
  - Politics is not all negative
  - Every aspect of evaluation is political—purpose and process alike
  - Evaluation reinforces but also disrupts power structures
  - Tension between evaluation, politics, power, and values is recurring
- Impact on the Evaluation
  - Understand evaluation's historical role in the program
  - Understand the relationships people and groups have with each other and with the program
  - Gain clarity about stakeholders' previous experience with evaluation

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

In thinking about the organizational, social, and political context for the RUPAS program, consider the aspects of the Family Matters (FM) organization that might impact the evaluation. What factors immediately stand out? What do you know about the community and how might it impact the program and the evaluation? What is the relationship of the RUPAS program to FM? What are the stakeholders' underlying motivations for commissioning or engaging an evaluation? If an evaluation was previously conducted, were the results used? If so, in what ways? What other organizational, social, or political aspects of the context might be relevant?

Resource

Google Maps
www.maps.google.com

As I mentioned, the "street view" feature here is quite useful and can be activated by going to the website above, entering the address of interest in the search bar, double-clicking the flag that marks the point of interest, and clicking the "street view" link.

Further Reading

Fitzpatrick, J. L. (2012). An introduction to context and its role in evaluation practice. *New Directions for Evaluation, 135*, 7–24.

The breadth of evaluation literature on context is reviewed in this article. Explicit analysis of stakeholder and program culture and how they influence evaluation are provided.

Vo, A. T., & Christie, C. A. (2015). Advancing research on evaluation through the study of context. *New Directions for Evaluation, 148*, 43–55.

A framework for understanding the various dimensions of context and how it can be used to systematically study context is outlined in this paper.

Weiss, C. H. (1993). Politics and evaluation: A reprise with mellower overtones. *American Journal of Evaluation, 14*(1), 107–109.

Carol Weiss offers a critical analysis of how and where politics enters and affects the conduct of evaluation. Much of what Carol notes in this paper still rings true. The article is a classic.

Quick Reads

1. Mary Crave on What's in Your Wallet? Or Back Pocket? Some Handy Questions for Encouraging Culturally Sensitive Evaluation
   http://tinyurl.com/zmtxc3j
2. Jennifer Greene on Context of Evaluation, and the Evaluator as Part of the Context
   http://tinyurl.com/jo5f5kf
3. Katherine Haugh, Smriti Bajracharya, and Kat Athanasiades on Putting Data in Context: Timelining for Evaluators
   http://tinyurl.com/gtflbqv
4. Mary Kane on Valuing Voice in Planning and Evaluation: Isn't It Obvious?
   http://tinyurl.com/j2h2be6

SECTION

How Do You "Understand" the Program?

What am I talking about in this section? *Understand* the program? Quite simply speaking, we want to know whether the things the program says it wants to do make sense. We accomplish this by addressing four seemingly straightforward questions: (1) What are the various elements of the program? (2) How do these elements relate to one another? (3) What is expected to happen when the program is implemented? and (4) Is there a logical flow between program elements and what is supposed to happen? Evaluators use a number of different approaches to organize answers to these questions. They all deal with creating visualizations—for example, concept maps, logic models, and systems maps.

The process of creating a concept map yields a visual web that summarizes how different *ideas* relate to one another. If "hunger" was a concept for which you were creating a map, for instance, you might write down any number of ideas that immediately come to mind. "Maslow's hierarchy of needs," "homelessness," and "developing countries" could be among them and they would be organized as spokes related to the root idea of "hunger." Other ideas might come up as well. The ability to document free-flow thinking is important in the process of creating a concept map.

A systems map, on the other hand, is a result of putting pen to paper to show the relationship between several organizations and processes. Using the "hunger" example again, we might be interested in

understanding how a government tries to address its food supply problem. This is a much larger idea than "hunger" alone. Rather, we need to consider who is a key player in the food supply chain. Perhaps farmers, farmworkers, truck drivers, and grocers; they play important roles in the food distribution process. Climate—though not a person—is also important to consider because it determines food quality and quantity. The availability of food also influences pricing. Likewise, policymakers, agricultural scientists, and corporations also need to be accounted for because they influence how food is grown and made. We can look to their decisions about farming practices, pricing, and tax infrastructures for evidence of their influence. Systems maps are helpful when trying to understand complexity and what happens to one aspect of the system when another one changes.

A logic model offers another means of displaying relationships—those that are more concrete than what is captured in a concept map and on a smaller scale than what is reflected in a systems map. Because we tend to conduct evaluations of individual programs, not abstract ideas or complex systems, logic models tend to be more manageable maps to work from and are the emphasis of this section.

LOGIC MODELS

A *logic model* is a depiction, or diagrammatic representation, of the various program activities and their linkages to program results. The underlying assumption is that a logical sequence exists between what a program does and what a participant is supposed to experience.

Now let us examine what a logic model might look like. A logic model is a picture showing the program's (1) inputs, (2) activities, (3) outputs, (4) outcomes, and (5) desired impact. Moreover, the logic model shows the specific linkages between them. Let me first define each of these terms.

- *Inputs* refer to the resources dedicated to the conduct of the program. This includes such things as money, facilities, community resources, and staff and volunteer time.
- *Activities* refer to what the program does with the inputs to achieve its desired impact. What things are supposed to take place within this program? What processes, events, and actions will be carried out in order to fulfill the purposes of the program? This includes things such as conducting a workshop, providing financial incentives, and having counseling or mentoring sessions.

- *Outputs* refer to the direct products of program activities. They can be tangibly counted immediately after the activities have been implemented and provide an indication of whether the activity was delivered to the appropriate targets and at an appropriate level. This means, for example, did the appropriate number of participants receive training? Or was the training of the appropriate duration? Was the training offered with great enough frequency?

- *Outcomes* refer to the benefits derived from having conducted the program. They are not as easily counted as outputs. Some outcomes can be measured in the short term, others in the medium term, and still others in the long term. Typically, we think of outcomes as benefits that program participants experience, such as changes in understanding, ability, or attitude. Furthermore, depending on the nature of the program activity and the related output, short-term outcomes might be measurable immediately after the activity is completed. For example, you might be able to gather evidence of early learning, or understanding, after a workshop or lecture ends. In contrast, you might need to wait a few weeks or several months to determine whether participants have truly been able to hone an ability—for example, to think critically or to drive with caution. Likewise, more difficult-to-measure outcomes, such as perception or attitude, might require a year or more to lapse before evidence can be gathered. Clearly, the program had expected participants to experience some set of outcomes. When thinking about outcomes, we consider the extent to which they were attained.

- *Impact* refers to the program's purpose or mission. In the long term, why is the program important? What marks does it ideally want to leave in the community where it is situated? On the people who participate in it? Impact is even more challenging to determine than outcomes—it is dependent on the logical flow between allocation of inputs (i.e., resources) to the design and implementation of activities and to the prior accomplishment of short- and intermediate-term outcomes.

Figure I.1 depicts the order in which inputs, activities, outputs, outcomes, and impact tends to appear on a logic model. The full sequence of inputs, activities, outputs, outcomes, and impact is the program logic model. Note that in this simple example, I have indicated two separate activities, each having its own output, and these outputs contribute to the accomplishment of three of the program's short-term outcomes. Output_1 leads to short-term outcome_1 and short-term outcome_2, whereas output_2 leads to short-term outcome_3. Short-term outcome_1 and $_2$, together, lead to the medium-term outcome_1 and short-term outcome_3 leads to medium-term outcome_2. Each medium-term outcome

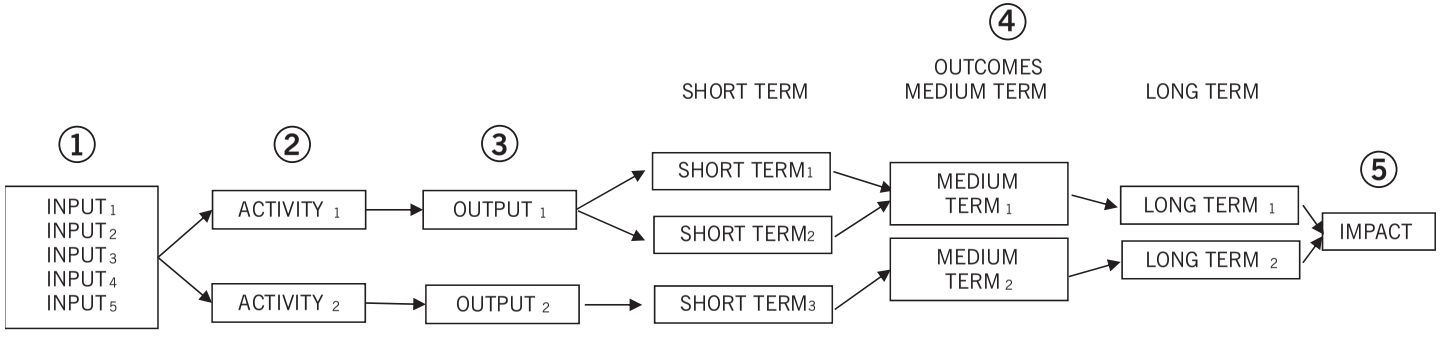

FIGURE I.1. Simplified logic model diagram.

leads to an individual long-term outcome and both long-term outcomes contribute to achievement of program impact. Any given element in the sequence results from the successful attainment of the previous element. Thus, in Figure I.1:

- *Inputs* (1) describe the resources provided for operating the program. These include such things as money, staff time, and facilities.
- Following this, if a program has these resources, they may be used to engage in particular *activities* (2). These include such things as training tutors, providing tutoring, and creating mentoring/advising relationships with students.
- If the program successfully accomplishes these activities, then for each they will attain an *output* (3) that designates that the appropriate amount or level of service was provided. This could include such things as the number of classes taught, number of hours of service delivered, and the number of participants served.
- In turn, if the program accomplishes a set of activities at the level intended, then participants will attain certain *outcomes*. (4) These could include such things as trained tutors, increased skills, and improved grade point average.
- Finally, if these benefits or outcomes are attained, then the organization's *impact* (5) will be accomplished. These might include such things as tutees' improved college readiness and tutors' increased orientation to community service.

An attempt should be made to understand why implementing a particular activity might be important for accomplishing a specified outcome or being able to engage in a subsequent activity. Hopefully, your gentle evaluator suggestions about these potential inconsistencies can lead to substantial program improvement at an early stage.

Figure I.1 is a fairly simple logic model diagram. Imagine, for example, a program that has many parallel activities, and further, activities that are contingent for their enactment on the accomplishment of a prior activity—that is, the attainment of a short-term outcome from an activity may be a prerequisite for the conduct of another activity. Or, several short-term outcomes together may lead to an interim-term outcome.

A further, more complete example of a logic model is shown in Figure I.2. This describes a UCLA tutoring program that employed undergraduate UCLA students to tutor middle and secondary school stu-

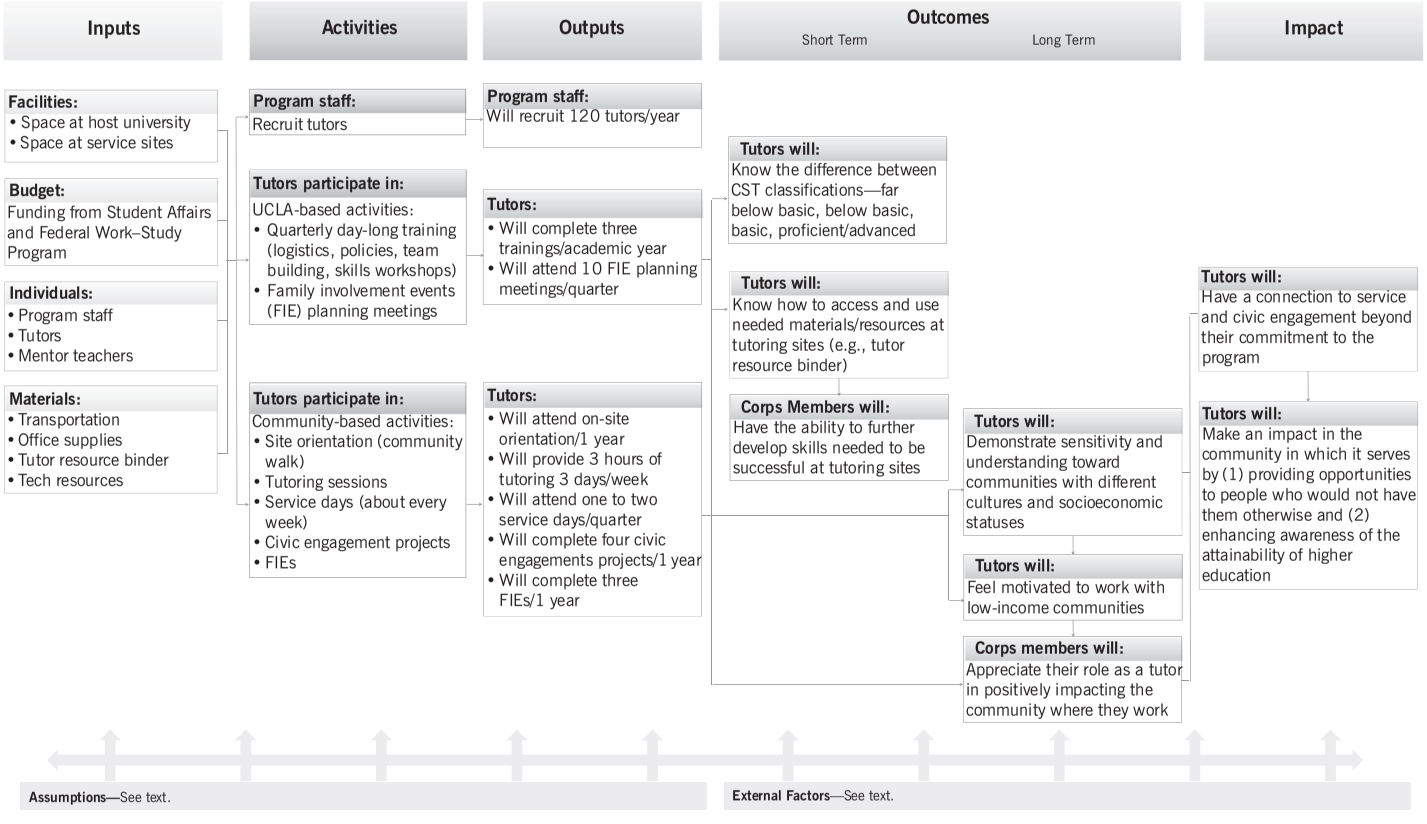

FIGURE I.2. UCLA tutor development program logic model—modified and simplified.

dents. The program was partially funded by a federal grant, along with resources provided by UCLA. In the model presented I have focused only on goals pertaining to the tutors. You'll note that each of the categories we have discussed is included in this modified and somewhat simplified model that my doctoral student evaluation team created in concert with program staff. Present in the figure are columns for inputs, activities, outputs, short-term outcomes, long-term outcomes, and impact. Note that two activities are shown in a stacked manner, indicating one was dependent upon the completion of the prior activity. For example, the recruitment of staff preceded their participation in goal setting. Similarly, several short-term outcomes together contributed to the attainment of a long-term outcome.

Observe the two boxes at the bottom of the figure titled "Assumptions" and "External Factors." I have not filled them in for this figure in order to reduce the complexity of the model. Under the "Assumptions" category are such things as the belief that tutors possess sufficient skill that they can be trained to provide tutoring. Of course, there is an assumption that tutors are also able to see connections between their tutoring and long-term community service and responsibility. Under "External Factors" we consider the kinds of events or actions that might forestall the completion of the process—that is, the continued funding of the program, mentors/teachers continuing support of the process, and availability of space that is conducive to learning. Moreover, there might be changes in the organizational context. While you would have considered the context in constructing the logic model, understand that contexts are always undergoing slight change.

WHY IS THIS IMPORTANT?

There are several reasons why developing and depicting a logic model is important. Obviously, the process of considering a program's intended sequence of activities and outcomes at an early stage can provide stakeholders and program personnel with an opportunity to consider the appropriate sequence of activities, how they relate to one another, what level of implementation (output) is to be considered appropriate, and what is the desired outcome. In general, it can lead to better understandings of the program. Sometimes the various staff members have different ideas about how a program should be run. Engaging in the process of examining a logic model helps the staff to create a shared vision or understanding of the program. It also provides an opportunity for staff to reflect on and ask questions of each other about their

work. Finally, it provides an opportunity for you to get stakeholders more vested in the evaluation as a precursor to potentially greater utilization of evaluation findings.

Furthermore, when program staff and other primary users examine the logic behind the relationships between activities, they may well decide to make important changes in the program. Too often, program developers simply list a set of activities that they presume will be important in attaining ultimate outcomes without reflecting on the reasons for those decisions. This might occur through having visited another program, seeing activities that were conducted at that site, and thinking, "That would be a good thing to do."

For the evaluator, the process of creating a logic model is an invaluable learning opportunity. The evaluator can learn about the program's evolution and the staff's history with the program as they reflect on their work. Much can also be learned about past and present challenges as well as successes. Of course, the model itself is important as well because it provides guidance about potential evaluation questions that might be considered. Since the logic model focuses on program activities, it allows the evaluator to pay attention to those activities and begin forming impressions as to whether appropriate implementation has taken place. The logic model will also direct the evaluator on the particular consequences (or outcomes) that might be associated with each activity and thus aid in more precisely selecting appropriate and relevant measurement instruments. It is also during conversations related to the outcomes reflected on the logic model that evaluators often learn of what data are already available versus what data must be collected anew, which has implications for how an evaluation is designed and conducted.

Often, evaluators need to assist program personnel and other users in developing and making explicit the logic model for their program. Three warnings are in order, however. First, coming up with a program's logical sequence is not really the evaluator's job. Ideally, programs will be so well constructed that activities have been carefully delineated and provide an appropriate rationale for the sequence of activities. In instances where that is so, it may not be necessary to engage in a logic model exercise—but having such a rationale is frequently not the case. The program's logical sequence might be implicit in the heads of program staff. They might believe that they have sensibly and systematically developed activities to fit certain purposes that would ultimately lead to the desired goals. However, unless their beliefs are made explicit, it is difficult to validate their assumptions about activities and their relationships with outcomes.

Second, I believe that it is important that evaluators not engage in the construction of a logic model based on *their own perceptions* of this kind of program. Developing a logic model *informed by* what research tends to say about programs of a similar type is part of the program staff's due diligence process. However, developing a logic model based *solely* on research is inappropriate. Evaluation, as I have noted innumerable times, should be focused on the particular program in question and the unique context that encompasses that program. Thus, the question really is, What did program developers intend to do and what was the logic behind that sequence of activities? Thus, your job is to assist program developers, program staff, and others in depicting the sequence of activities leading to the attainment of outcomes.

Finally, do not think that the job is done once a logic model is finished, and don't let program staff think that a logic model is a static document. Every aspect of a program is potentially subject to change. Thus, every aspect of the logic model is open to modification. The social nature of programs makes them dynamic entities. Availability of funding fluctuates. Staff come and go. The need for the program waxes and wanes according to the community's needs. The point is that program staff have creative freedom to adjust the program as they learn about what works in their particular environment. The logic model should be viewed as a living document and should be adjusted to reflect these changes.

GETTING STARTED

How do you get started? The notion of a program's logic or of logic models is not usually well understood by potential stakeholders. A good starting point is meeting with stakeholders to explain what a logic model is and why it is important (much as we have discussed earlier in this section). It is important for you to point out that while the exercise will benefit you, as the evaluator, in being better able to measure processes and outcomes, it also is of great value to them and the program. Furthermore, the process of developing the logic model will help stakeholders to develop a common language to describe what it is that they are doing and will facilitate greater understanding of their program.

At UCLA, I currently have an evaluation capacity-building project funded by the university. It focuses on the dozen or so university programs designed to provide academic support to underrepresented students attending various elementary, secondary, and postsecondary

schools in the Greater Los Angeles Basin. My doctoral students and I work with staff members from these projects to initially develop logic models that describe the program activities in which they are engaged.

What do we do? We typically start with a several-hour workshop where we discuss what logic models are and describe what they look like. The next step is to have participants bring written materials about their program as references when constructing their own logic models. Initial logic models, if they have one, are also brought to the meeting. Participants are asked to answer four questions:

1. What impact does your program hope to have?
2. What must your program accomplish to attain this impact (or what outcomes do you think contribute to your impact)?
3. What is the list of activities that are a part of this program?
4. At what level must these activities be accomplished?

This discussion is stimulated by a consideration of subquestions that ask what the program consists of, what the program does, and what sorts of services are provided.

You will notice that the first of these questions relates to impact, as I have described them earlier, the second relates to outcomes, the third describes the activities that are intended to take place, and the fourth refers to outputs. Missing from this question list are the inputs (or resources) that are to be provided. We tend to put this on hold temporarily until much of the rest of the model is worked out.

Participants produce a long list of items for each of the categories. We work with them to examine the list to determine closely related items that might be grouped or combined. We also try to reduce the lists by combining repeated themes or ideas.

The next step involves listing the activities, outputs resulting from those activities, short-term and long-term outcomes, and impact in separate columns and discussing the way in which items in each column relate to the next. For example, what specific outputs are expected through engaging in a specified activity, and, moreover, how might the activities at that output level contribute to an outcome?

We do this by transferring items in the lists onto sticky notes, with different colors for activities, outputs, outcomes at different levels, and the impact. This makes it easier to arrange them into sequences, chains, or pathways. We draw arrows to show these relationships and to indicate a probable direction of influence and point out that sometimes activities are sequentially related to each other. For example, program

staff recruitment precedes and enables staff training, while training enables the implementation of an activity. Thus, arrows can be drawn between these stacked activity items in order to show these relationships.

In the working session, we provide four pieces of advice: (1) arrows should show the relationships between individual items in the logic model, (2) every activity should have at least one arrow leaving it, (3) every output should have one arrow entering and at least one leaving it, and (4) every outcome should have at least one arrow entering it and may have an arrow exiting it to another activity or to a longer-term outcome.

A dilemma that evaluators sometimes face is where to start. Does one start with the ultimate goals and indicate what kind and sequence of outcomes would be necessary for their accomplishment? And, in turn, what activities at what output levels lead to those outcomes? Or does one logically proceed from activity to output to outcome? That is, if programs plan to implement particular activities, what potential outputs do they anticipate from each, and how does that lead to desired outcomes? As you can see, from the description of the workshop that we conduct, the process is allowed to work in both directions simultaneously.

With respect to inputs, a topic that I avoided earlier, it is important at some point to specify the nature of those resources that are being made available. Assuming that one was evaluating a continuing program, then the inputs generally were thought to be sufficient for providing the activities, which were specified. Consideration of the inputs reaches greater importance if in the discussion of the logic model discrepancies in relationships arise—for example, if an activity does not logically lead to a needed outcome. This might suggest the need for a modification of activities. In that instance, the inputs that are available provide the resource boundaries for what activities might be added or modified.

So, actively engaging primary stakeholders in the process of constructing a logic model provides a depiction of the program logic—the logic behind what the program intends to do. This depiction is a representation of the program at one point in time—but only one point. Logic models are "living documents" that can and will change over time and should be periodically revisited with stakeholders.

RECAP—SECTION I

Understanding the Program's Logic

- The Need for Understanding the Program's Logic or Rationale
- Logic Models
  - Elements of a logic model
    - Inputs
    - Activities
    - Outputs
    - Outcomes
    - Impact
- A Partial Logic Model (Example)
  - Stacked activities or outcomes
  - Multiple short-term leading to intermediate-term outcomes
  - Assumptions
  - External factors
- Why Is This Important?
  - Program staff can benefit
  - Evaluator's job made more explicit
- Getting Started
  - Explaining concepts to participants
  - Bringing in relevant materials
  - Identifying activities, outputs, outcomes, and goals
  - Understanding relationships
  - Are the inputs sufficient to accomplish the outputs and outcomes?

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

We discussed the process of creating logic models in this section. Of course, to the extent possible, you will want to engage program stakeholders in this process. Before doing so, however, it will be important to consider these questions: To what extent are stakeholders familiar with logic models, the logic modeling process, and its purpose? What do you want them to take away from the logic modeling experience?

As you are creating the logic model, note some of the activities that take place in the RUPAS program: creating booklets, identifying parent leaders, organizing parents in a community, organizing community meetings, and supervision by RUPAS staff. Are these all of the activities? What else would you add? What do you believe to be the short-term outcomes related to activities or groupings of activities? What do you think is the goal or expected long-term outcome of the program? How do your perceptions compare with those of stakeholders'? Specifically, consider their assumptions about the program and the relationships between activities and outcomes.

Let's now take a step back from the possible "doing" and pause for a moment. Reflect on the ways in which the process of creating a logic model in partnership with stakeholders could facilitate discussions about their conceptions of and assumptions about the program. Why do you think it would be important to explore these assumptions? Also, whose views and which values are represented in this discussion? If you were to create a logic model for RUPAS, what might it look like?

Further Reading

Centers for Disease Control and Prevention. (2016). Logic models. Retrieved from *www.cdc.gov/eval/resources/index.htm*.

This site provides introductory materials on the logic model and how you might go about developing one.

McLaughlin, J. A., & Gretchen, B. J. (2015). Using logic models. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of practical program evaluation* (pp. 62–87). San Francisco: Jossey-Bass.

This chapter describes a logic model process in great detail. The authors stress that the logic model is a "useful advance organizer" for conducting an evaluation.

W. K. Kellogg Foundation. (2006). Logic model development guide. Retrieved from *www.wkkf.org/resource-directory/resource/2006/02/wk-kellogg-foundation-logic-model-development-guide*.

Similar to materials from the CDC website, the Kellogg Foundation also offers reference materials for better understanding logic models, how they can be created, and subsequently used.

Quick Reads

1. Ann Gillard on Using Logic Models
   http://tinyurl.com/zq82u9a

2. Tara Gregory on Using Storytelling to Help Organizations Develop Logic Models
   http://tinyurl.com/hf9j443

SECTION

J

What Are the Questions/Issues to Be Addressed?

The evaluation questions form the basis for an evaluator knowing what to do in the evaluation. They are an essential part of subsequently developing an evaluation plan. But how do we get to that stage? Developing an understanding of the program is a necessary first step for the evaluator. Next, examining or helping to develop a program logic model (as we have discussed) is often useful in developing the evaluation questions—or more broadly, identifying and organizing the issues to be examined. But there are many aspects to a program and many potential questions that might get raised. It is simply not possible to look at everything that key stakeholders might state as questions of interest. There are limits. So, what might you do? First, you need to learn and take stock. Learn why program staff are seeking out an evaluation now. Take stock of the kinds of questions that can be asked and answered during an evaluation. Then, engage stakeholders in defining and refining the focal evaluation questions.

WHAT MOTIVATES EVALUATION?

Recall that in Section H we discussed the importance of context. That topic returns in this section. As evaluators, we must understand what is giving rise to the need for an evaluation in this particular program, at this point in time. Evaluators often hope that evaluation is a natural

part of every program's everyday operations, but that rarely is the case, however, simply because time and resources are finite. Still, there are a handful of possible scenarios that can lead to a program commissioning an evaluation.

For instance, program staff sometimes take the initiative to engage in evaluation for the purpose of satisfying their own curiosity or to improve what they are doing. They might also find themselves engaging in evaluation to satisfy funding requirements. What do I mean by that? Many programs are externally funded—sometimes from larger governmental agencies or nonprofit organizations such as foundations. For example, a local welfare program might receive funds from a state agency for a particular local program. In these instances, the funding agency may have specific reporting requirements. These requirements may simply focus on a verification of the activities implemented and the services provided. These kinds of process measures, therefore, need to be part of the evaluation agenda. It's also possible that the agency could demand that certain outcome data be presented—this must be complied with.

In other instances, a program might be funded and approved by its own agency, but at a higher organizational level. A school district, for example, might approve allocation of resources to develop a weekend tutoring program at the school level. The staff at the school had heard about the new approach to teaching mathematics and asked to modify an existing program. In this case, the district may have provided additional resources for this "new" program. Even if no new resources had been provided, the act of authorizing the conduct of the program carries with it specific obligations. What is it that the school superintendent expects will happen as a result of this program modification? What does the staff want to see? What are the implicit (or explicit) questions that emanate from that level? Further investigation might be necessary.

WHAT KINDS OF EVALUATION QUESTIONS CAN BE ASKED?

We have already asked a number of questions before even getting into the thick of the kinds of evaluation questions that stakeholders might want answered. Surely, program staff might simply ask, "What are we doing?" (or "How are we doing?"). Alternatively, their questions could be centered on whether their program is reaching its goals (or achieving its outcomes). These are three very different questions, each focusing on three very different program elements. Recall our logic model discussion in Section I. A logic model has five major components:

(1) inputs, (2) activities, (3) outputs, (4) outcomes, and (5) impact. Do you see any parallels between these questions and components of the logic model?

Let's think about it this way. Which components of the logic model might you reference if you wanted to get at "What are we doing?" Perhaps inputs and activities. Why might this be? At the heart of "What are we doing?" is the desire to know about what programmatic decisions are being made and how program activities are being implemented. It is a question about program processes (more on this in Section O)—that is, how the program got started and got off the ground. Let's think about this a bit more. What, exactly, could program staff ask about program processes and activities? One issue is whether the stipulated program activities are being put into action in the manner expected. Examples of such evaluation questions might include "To what extent are they being implemented as intended?" (or "Are program activities being implemented with fidelity?"). Various questions of this type might be generated. In the early stages of program implementation, the concern might relate to the way in which various activities are interacting with one another. Are there unneeded redundancies? Was the logic that led to the creation of these activities faulty? Program developers might have made judgments about the manner in which participants would engage in the activities and their reaction to them. Did these judgments bear out? Some other issues might be about participant satisfaction or interest.

What if staff wanted the answer to "How are we doing?" Which segment of the logic model might you examine? A close look at outputs could help. You might ask, "Why? That's such a simple question. Shouldn't we just know?" You are *partly* correct. "How are we doing?" is a seemingly simple question, but it is complex. We need to know what "How are we doing?" means. "Doing," with respect to what and when? Remember—outputs have to do with counting those things that are tangible immediately after implementing a program activity. Phrasing the question as "How are we doing?" does not trigger our minds to think about enumeration or activities. Thus, we must reframe and ask if staff wants to know: "How many activities are offered? How frequently are they offered? How many participants are completing program activities? How do these figures compare with last month versus 6 months ago versus a year ago?" Answers to these kinds of questions allow staff to accomplish two things: (1) monitor performance within the program and (2) ask questions about whether the frequency, rate, or dosage at which activities are being offered contribute to achievement of program outcomes. It also provides opportunity to wonder whether some activities should be streamlined if there are several taking place

at the same time and for the same purpose. If there aren't very many activities going on, should existing activities be delivered with greater intensity or frequency? What other programmatic adjustments might be made to achieve the program's goals?

Speaking of goals, which part(s) of the logic model could shed light on goal achievement? We have already examined inputs, activities, and outputs. By process of elimination, we would most likely look at outcomes and impact to answer this last question. If you recall our simplified logic model in Section I (i.e., Figure I.1), outcomes can be divided as short, medium, and long term. Which ones are we to look at? All of them? No, not likely, and for several reasons. There aren't enough resources and time to examine everything. Also, the program may not have been operating long enough for us to reasonably expect it to achieve its long-term outcomes. Lofty goals in particular (e.g., curing an addiction to smoking, developing more responsible citizens) often cannot be accurately determined in the short periods of time typically allotted to an evaluation. The program may very well have to continue for a longer period of time before such outcomes can be seen. Thus, short- and medium-term outcomes offer well-reasoned starting points for identifying evaluation questions about program goals.

You might wonder, at this point, if evaluation questions about program outcomes are as simple as "Is the program reaching its short- and medium-term goals?" That is certainly a first step in the right direction. You might even rephrase that question as "What kind of effects does the program have on its participants?" In the case of the UCLA tutoring program discussed in Section I, the answers to whether the program is reaching its goals could be as clear as "yes" or "no," or it could involve a bit more unpacking. You might conclude, at the end of an evaluation, for example, that "The program is doing well teaching tutors the difference between various California Standards Test classifications (short-term outcome_1) and motivating tutors to work in low-income communities (medium-term outcome_2). However, it is not as effective in helping tutors learn how to access and use materials (short-term outcome_2), or teaching them how to be culturally sensitive (medium-term outcome_1)."

Instances where the answer isn't so clear suggest that there are opportunities to probe deeper into the nature of program outcomes themselves. Continuing with the same UCLA tutoring program example as above, you might ask, "Are we seeing any evidence of tutors developing skills to be successful at their school sites (short-term outcome_3)? Is it taking longer than program staff expected? Perhaps it is more appropriate to move that to the medium- or long-term column?"

Likewise, you might notice that tutors becoming more culturally sensitive (medium-term outcome$_2$) has less to do with their motivation to work in communities of need (medium-term outcome$_3$) than their compassion for people from different backgrounds in general. This outcome is not on the logic model. You might suggest that program staff revise their model accordingly. Also, you might consider the linkages (or the arrows) between various program elements more carefully. This type of thinking leads to the formulation of questions about causal mechanisms (we address this in more detail in Section P)—that is, does delivering a program activity at the current frequency contribute to attainment of short-term outcomes? Similarly, which short-term outcomes contribute to achievement of medium-term outcomes, and so on. Thus, the focus (or one of the foci) of the evaluation might be to examine a simple relationship within the logic model.

You see, no matter which questions program staff might have, you should be able to map them back to the logic model. Likewise, the logic model could inspire the identification and formulation of questions. This is precisely why creating a logic model at the start of the evaluation is helpful—to staff and to you, the evaluator. Now, what if you don't have a logic model to work from because you were unable to create one with program staff? How might you go about articulating evaluation questions then? There are other materials and resources that will be helpful. We discussed programs being funded by outside organizations earlier. When that is the case, we can reasonably expect that program staff must have gone through the process of applying for a grant. If you have not already done so, ask the staff to share a copy of the call for proposals (or request for proposals) along with their grant application. These documents, together, should give you a fairly accurate view of what the funding agency requires for an evaluation, what program staff proposed to demonstrate as evidence of success, and how you might be able to guide discussions about the evaluation questions moving forward.

GETTING STARTED ON DEFINING QUESTIONS

Clearly, there is a world of possible questions that might get examined. What should you, as the evaluator, consider in eliciting the questions or issues that key stakeholders believe to be important? Let me suggest some factors that might guide your evaluation efforts:

1. Communicating the evaluator's role.
2. Identifying the stakeholders to be involved.

3. Emphasizing the desire to pursue meaningful and useful questions.
4. Finding questions that need answers.

The Evaluator's Role

First, a cautionary note: It is important that you recognize that the decision about evaluation questions is not yours to make. You, as the evaluator, want key stakeholders to own the evaluation. It is important that the stakeholders come to recognize that the evaluation is being conducted to suit their program's needs. You might be aware of issues raised in the literature related to programs of the type being evaluated. However, as I noted in the prior section, the primary focus of questions or issues to be addressed should not be dominated by the research literature. Programs cited in those writings may very well be different from the one you are evaluating—and the context most certainly differs. Relevance for local needs should be the first priority. Your interest in hearing their questions, and focusing on their questions, should be strongly affirmed at the outset. Make clear that the questions to be examined will be their questions, based on their concerns. These questions may be subject to later revision.

Stakeholders to Be Involved

One of the first tasks in determining the questions or issues to be addressed is identifying the stakeholders who are to be involved in this process. I have considered this question generally in Section E, but it is so important that a very brief review is in order. I have noted that there is a broad group of individuals who might be considered stakeholders in a program. Furthermore, it is my view that the evaluator cannot attempt to deal fully with all of these groups on major decisions. Deciding on the evaluation questions is that kind of major decision. This issue is very important since the evaluation questions chosen will shape the evaluation. There is a group—a manageable group—of involved stakeholders whom I have referred to as primary stakeholders. You, as the evaluator, want to know what questions are of specific interest—maybe even urgent interest—to these involved stakeholders.

While these primary stakeholders are the individuals who will be proposing potential evaluation questions, you realize that for the evaluation to be of value it is necessary to understand other points of view. To put it simply, their view is not the only view in town. The views of the broader stakeholder group should be heard. But let's not

get confused; it is not possible to satisfy everyone's desire for the evaluation, so do not enter an evaluation with this unreachable goal in mind. However, it is helpful for you to "hear out" what those alternate views might be. You might be working with a program director and some key staff, but you, as the evaluator, should meet with others. Put your ear to the ground, so to speak. What concerns do participants in the program have? What about program staff? What about others in the community who might be impacted by the program? What are the views of marginalized groups that may not have the opportunity to actively participate?

Your obligation as the evaluator is to alert the primary stakeholders—who are your major decision-making constituency—about the other views and issues that might have relevancy. Your task is not to impose these questions from other groups or to imply that they should be included. You simply want to alert those who will be selecting questions of evaluation concern about the additional possibilities.

Pursuing Meaningful and Useful Questions

Unlike research studies where the investigator often frames the questions, evaluations are owned by stakeholders. You want to be able to answer questions that are important to them. Thus, I would like to discuss the process of working with primary stakeholders to develop the evaluation questions.

First, do these stakeholders have concerns? Do they want to know something about the program? What are they unsure about? Initially, you might set no limits on this inquiry—any question is fair game. Participant stakeholders should be made to feel free to say whatever they think is important. Encourage them to say whatever comes to mind—no restrictions. You, as the evaluator, should encourage this openness to the presentation of a broad spectrum of possible questions. There will be ample opportunity to sort through, reflect, and modify these questions at a subsequent time.

If participants have difficulty getting started, the evaluator might offer some general assistance. For example: Do you want to know something about the program's outcomes? If so, which ones? Or are there some aspects of the program and its operation that concern you? Also, what kind of decisions might the evaluation help you to make? Do you want to judge the program as an entity? Do you want to see where to make changes? These types of questions should spark some kind of response. What is important to remember is that general guidelines, as above, should remain general. Do not suggest specific questions, and do not let your disciplinary inclinations "lead" the evaluation.

At some point there needs to be a narrowing of the best questions. Now you and primary stakeholders need to examine these sorts of questions and refine and consider them. You, as the evaluator, should seek to assist stakeholders in clarifying the specific tentative questions raised. You might ask for clarification. "I'm having trouble understanding what that means, can you restate it in a simpler form?" This is a first step in clarifying questions. The question should be perfectly clear to you, or it is not possible to proceed. If you don't understand it, then you can't evaluate it.

Furthermore, you need to determine if the questions are definable and specific enough. For example, the question, "Is the program doing well?" is not answerable in that form. What does "doing well" mean? You need to get to specific questions that clearly delineate what is considered good, so that "doing well" will have meaning.

Finding Questions That Need Answering

Another step in the specification of meaningful questions is determining whether the questions are ones that really demand an answer. What do I mean by "really demand an answer"? You want the evaluation to help answer a question, so the issue, then, is whether the question is real. You want to know whether the answer is already known. If stakeholders already know the answer, why ask the question? Is it relevant? You want to be assured that the question is not so trivial that it's not worth pursuing that issue.

Some have attributed Einstein as saying, "If I had only one hour to save the world, I would spend fifty-five minutes defining the problem, and only five minutes finding the solution." This is a particularly poignant way of emphasizing the importance of understanding stakeholders' information needs—that is, what questions they need answered and the extent to which those questions are of high quality. Questions might be brilliantly stated but if people don't want answers, then they are not of value. You need to know, for example, whether a course of action has already been decided upon and this evaluation question is just window dressing. Have they already decided what they believe is the answer and the action that will be taken? Is there no, or little, likelihood that the answers to the evaluation question will lead to program changes or even to changes in attitudes, understanding, behavior, and the like?

In my work, I like to apply a litmus test of sorts. I might take the question that they have stated and propose some realistic potential evaluation findings. I develop some "what if's." Then, I would inquire as to what that finding would mean for their program. I might ask

if there were changes they might likely make in the program based partly on these possible findings. The issue of whether the proposed questions are those that primary stakeholders really want answered is further pushed by varying the findings and asking yet again what future actions might be implied. For example, I might present the most positive findings and ask the questions of possible use. What difference would it make? What would you do differently? Would it change your current opinions? And then, again, I might provide potential negative findings and ask about the implications. In instances where some kind of participant response would appear to be warranted but no action or change of attitudes is elicited, I would inquire further. I might ask why they feel that no action is warranted. I might say, "If we do the evaluation and the outcomes are positive (or negative) and you then see no action would be taken or views changed, why do you want to ask this question?" Typical responses might be that the evaluation findings were not on target with what they had in mind for the evaluation question. This either leads to a further refinement of the question or to a better understanding on my part about appropriate measures for the question.

Sometimes the response to my question about potential use elicits a different kind of statement, which is also quite fruitful in focusing on the proper evaluation questions. This is a kind of contingency response—"Yes, but." Let me explain this by providing an illustration. I might ask stakeholders to imagine, in evaluating a school mathematics program, for example, that there were positive findings on the mathematics test used to measure the program's achievement. So, I then prompt them with, "Would you plan to extend the program or to continue the program?" You might receive the response, "It depends." Depends on what, you might ask. "Well," the involved stakeholders might say, "on whether students enjoyed or didn't come to dislike mathematics, and on whether teachers were positive about the program." Thus, in this line of questioning you have potentially introduced additional questions important to stakeholders that might have real relevance. The mathematics achievement by itself is not sufficiently important to lead to potential action, but use of the evaluation results is also dependent on student and teacher views. These issues may very well merit having evaluation questions. This too should be discussed.

You want to conduct a meaningful and useful evaluation. Let's face it, life is too short to engage in evaluations that don't have meaning to stakeholders and are not likely to be used. Focusing on important, needed, and wanted questions is the first step. The bottom line is to engage stakeholders during the process of suggesting questions in a manner that helps to determine whether they really want answers. Are

they potentially amenable to inspiration, and the possibility of change? Does someone intend to use the evaluation?

RECAP—SECTION J

Getting Started on Defining Questions

- Evaluator's Role—Working toward Stakeholder Ownership
- Identifying Stakeholders to Be Involved
- Pursuing Meaningful and Useful Questions
- Finding Questions That Need Answering
- Apply the "Litmus Test" of Potential Use

➢ ***Some Next Steps:*** Getting started in identifying *tentative* questions or issues is really just a first step. Before finalizing the evaluation questions, I ask you to consider several other things. In Section M, we examine sources from which data are acquired for answering questions, and in Sections K and L, I present a description of the methods used for acquiring quantitative and qualitative data, respectively. Finally, in Section N, we reexamine whether the questions are answerable. In doing so, I ask you to consider some additional insights: Are the evaluation resources sufficient to answer the question? Are the questions stated in a way that information could be attained? Are there potentially appropriate instruments available? And so, *read on*.

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

Now that you have developed a draft logic model for the RUPAS program, let's use it to guide our planning and thinking about the evaluation that we are trying to conduct. Let's think, for instance, about the questions that stakeholders would like to have answered. What do stakeholders *want* versus *need* to know?—that is, if you involved *only* Amy Wilson in the logic modeling process, on which area of the model do you think she would want to focus when developing evaluation questions—inputs, activities, outputs, outcomes, or other? What if you had engaged only Children's Trust? Or only the parent leaders? How might the focus on questions and the actual questions to be answered change? Note these questions, as they will be important in subsequent case exercises.

As you consider the nature of these questions, I also encourage you to think about why each of the stakeholders (or groups of stakeholders) involved desire answers to these particular questions. What is their motivation? Do they comple-

ment or compete with each other? Furthermore, because it is never too early to think about the potential use of evaluation findings, consider the courses of action that they intend to pursue based on what they learn. Are stakeholders prepared to act on findings?

Further Reading

Cooksy, L. J., & Mark, M. M. (2012). Influences on evaluation quality. *American Journal of Evaluation, 33*(1), 79–84.

This article stems from Leslie Cooksy's 2010 AEA Presidential Address in which she asserts that at least three factors affect the quality of an evaluation—an evaluator's competency, the nature of evaluation context, and the nature of resources available for evaluation. Evaluation questions are embedded within these factors, thus, this paper offers one lens through which you can consider this issue.

Friedman, V. J. (2006). The power of why: Engaging the goal paradox in program evaluation. *American Journal of Evaluation, 27*(2), 201–218.

This author indicates a procedure for examining how to evaluate which goals are most important. This might be helpful in thinking about the issues of focusing on important questions in an evaluation.

Quick Reads

1. John Cosgrove on Questions and People Drive Continuous Improvement
   http://tinyurl.com/jkbdk8c
2. Shana Alford on the Art of Asking the Right Evaluation Questions
   http://tinyurl.com/zba4ywr
3. Leslie Goodyear on the Importance of Asking "Stupid Questions" in Qualitative Evaluation
   http://tinyurl.com/z275cyq

SECTION

K

What Are Instruments for Collecting Quantitative Data?

In this section, I discuss the instruments frequently used to collect quantitative data. Any procedure or instrument that provides information that can be quantified—meaning numbers can be attached to the information—is a quantitative method. Quantitative methods show the degree or extent to which certain characteristics are present. Why might you want to collect quantitative data? Quantitative data are in many ways easier for people to understand and grasp. Numbers seem to be concrete. When one sees a number, there is the concept of value, size, and degree. Furthermore, comparability to other numbers is more easily understood. For example, if you examine a student's score on a math test, you can see how well he or she understood the material based on the score received. You might do that by looking at the accomplishment of the student in relation to other students in the same class, or to those outside of that class. If you look at the test scores for all of the students in the class, you can identify what most students understand and areas where they struggle. If the teacher has expectations about student progress, then the test score can provide an indication of whether those expectations have been met. Such a test is just one example of an instrument used to collect quantitative data—there are many others. In this section, I talk about the instruments (means of collecting data) that *most typically* yield *quantitative* data.

TYPES OF INSTRUMENTS FOR COLLECTING QUANTITATIVE DATA

The quantitative data obtained in evaluation may be related to aptitude, knowledge acquired, skill attained, attitudes, or values. Several types of instruments or measures that individuals respond to can be used to obtain this data. *Tests,* as I mentioned above, are often used to gather information about potential ability, learning, and skills. Data about skills, especially when thought about in terms of behaviors, can be captured using *logs* and *checklists*. *Surveys* or *questionnaires* can be used to measure attitudes or perception. Let's look at each of these more closely.

Tests

Tests are measures designed to determine whether knowledge was acquired or skills were attained. There are several examples of very well-known tests—most of them are standardized and administered on the national level. The Graduate Record Examination (GRE), for example, is an *aptitude* test used for determining competitiveness for graduate school admittance. This is an instrument that presumably measures potential ability. In evaluation, we very infrequently are concerned with potential ability as an outcome measure. Programs are not designed to improve aptitude, but to measure what results actually came out of the program—things like learning, attainment, or changes in attitudes. Ability tests, however, might be part of an evaluation in instances where we want to examine the initial comparability of the program participants and of a comparison group. Also, perhaps an aptitude measure might be administered at the beginning of a program for the purpose of participant selection. In such instances, a key question might relate to the impact of the program on groups of students who exhibit different aptitude levels.

More frequently, the tests used in an evaluation are concerned with *knowledge acquisition* or *learning*—that is, did students learn seventh-grade mathematics or achieve at a seventh-grade level in mathematics? Or did they learn the causes of a disease? Tests can also be used to determine one's skills. For example, did participants learn the steps required to assemble a complex set of machinery? Or can students accurately diagnose an illness and prescribe the best treatment plan?

Logs and Checklists

Quantitative data may also be acquired using other instruments. One such example is a *log* (or structured diary). Information gathered

through these instruments tends to be self-reported and is collected on an ongoing basis in a prescribed manner. A daily log might ask individuals to supply quantitative data on various behaviors or activities in which they engage. Consider, for example, such things as the number of cigarettes smoked; the amount of time spent on homework or watching television; or, perhaps, the quantity of various foods eaten. Information concerning when each activity occurred and how much time was spent on each activity must be documented.

Checklists offer another avenue by which quantitative data about skills, and behavior more specifically, can be collected. In the case of logs and diaries, program participants self-report data. Checklists are often completed by the evaluator or someone on the evaluation team who has been trained to use the instrument. Data collected using checklists are done through observation. Normally, observational data are qualitative—and I talk more about that in Section L. However, observational systems can be devised that provide quantitative indications of what is occurring within a program setting.

In using such protocols as logs or checklists, behaviors and activities must be well defined. Observers should be given detailed instructions for identifying the behavior(s) of interest. Preferably, a training session should occur to verify that the observer(s) can properly identify what is to be looked for. Furthermore, data collected in this manner should be done at well-specified times and for the same durations. In an instructional program, for example, the amount of time the participants are attentive to the classroom task can be recorded using a log. If there are multiple activities to be observed within a program, an observer might have a list of those particular aspects to be observed and can record either the occurrence of the event or the amount of time spent on each.

Surveys

Another instrument that yields quantitative data is a *survey*. Surveys employ procedures in which people are asked a series of questions—the answers to these questions typically yield quantitative data. Surveys can be conducted by telephone. (Think of the calls that you have received after a visit from the cable company or being on hold with your cell phone provider requesting your feedback about the quality of service delivered.) Surveys may also be conducted verbally, perhaps by someone waiting outside a grocery store to ask you a few questions.

I like to consider a *questionnaire* as a particular type of survey—one that is administered in written form. We are most familiar with paper-

and-pencil questionnaires. Undoubtedly, you have filled out many questionnaires asking about your attitudes toward something, or satisfaction about a program, or your evaluation of an instructor. These days, modern technology has allowed the administration of questionnaires through the Internet. I am referring to this innovation also as a questionnaire because responders, themselves, are, in fact, reading the questions and providing a written response.

FINDING EXISTING INSTRUMENTS VERSUS DEVELOPING NEW ONES

Quantitative data can be found anywhere. Data are quantitative if one contrives a way of enumerating the information. This type of data might be obtained and available to you even before you begin an evaluation. This *existing data* may be found both within the program and outside of it. Within the program there are certainly available things such as records of attendance, absences, health records, or referrals to other agencies or programs. There may be test data that have been previously collected. External to the program, but also existing, you might find valuable data in various governmental records—possibly census data, crime statistics for the community, or economic indicators. Again, as I have stated numerous times, you, as the evaluator, must be guided by the questions being asked in determining the relevance of existing data. It is tempting, of course, to think that a portion of your data collection job can be fulfilled with the ease of using existing data. However, *select only what is needed and what is relevant to the questions being asked*.

More frequently, however, you will need to collect *new data* because relevant existing data are unavailable or insufficient. New data might be acquired by administering an *existing instrument*—that is, instruments that you can access and use—perhaps, for example, existing published tests. Furthermore, if existing instruments that meet your needs are unavailable, you may need to create a *new instrument*. We discuss each of these courses of action in turn.

Finding Existing Instruments

How do you locate existing instruments? The evaluation question that you are attempting to answer might have specified certain kinds of attainment on a particular instrument, or test, as its objective. For example, you might be asked to demonstrate through an evaluation

that students' scores increased on a particular test (e.g., a state assessment test). In that case, there is no problem—the choice of instrument is clear.

What to do if that is not the case? One possibility is that tests or other instruments were a part of the program's instructional materials. Are these valuable for answering your evaluation question? The answer to that depends on the nature of the question. If the question is solely related to the prescribed curriculum of the program, the test may suffice. If, on the other hand, the question goes beyond that, then other instruments are necessary.

There are books that summarize tests that are available. One such series of books is entitled The Mental Measurements Yearbooks, which are updated every 2 years. You could look at psychological abstracts based on a topic of interest to see what tests or instruments were used in previously conducted studies that might satisfy your unique question. Also, major test publishers have catalogues listing their test instruments for purchase. For example, the Educational Testing Service has a *Test Collection Catalogue*. You might inquire about these at the reference desk of major libraries. Finally, thanks to modern technology, try going to Google or a comparable search engine and indicate your topic (e.g., "measuring social adjustment"). You will be surprised at how much information you are able to find.

It is sometimes easiest to use an existing instrument, particularly when the question is related to measuring achievement. There are many such tests available that have been standardized. By "standardized," I mean the test has been developed, administered, and scored in a consistent manner. In essence, this means that there was special care taken in selecting test items, the conditions for administering the instrument have been well specified, and the way in which interpretations are made has also been indicated. All of this is documented by the test publisher. Perhaps the most important aspect of a standardized test is that it has been "normed"—that is, the test had been administered to a large group of test takers and the results of the procedure provided the basis for potential comparisons. The test takers constituting the norming group, in essence, are a comparison group. Scores that those being evaluated achieve can then be compared with this norm. In essence, the test taker's relative position can be determined. Not surprisingly, tests of this type are called *norm-referenced tests*.

There is a caution to be sounded at this point. You must examine who was included in the norming population to determine whether they are truly comparable to the program participants you plan to test.

If they are not comparable to your population, then comparisons might be meaningless.

One aspect of the standardization that occurs in norm-referenced tests is examining the reliability and validity of the instrument. *Reliability?* Think about it. What does "reliable" mean? If someone is reliable, you know that you can depend on him or her—the person is consistent. Reliability is measuring consistency. If you give the test multiple times to the same person, it will yield consistent results. Assuming that the conditions of administration were the same, the score will not fluctuate—or if so, not by much. Standardized, norm-referenced tests will report reliability coefficients. This is helpful information in determining whether you want to use the test.

A test might be reliable but that does not mean that it is necessarily valid. The dictionary describes the word "valid" as relevant, meaningful, and appropriate to the end in view. In essence, *validity* refers to the degree to which the test, and data from it, appropriately captures the concept that the test purports to measure. For example, does the mathematics test really capture the essence of mathematics for the grade level in which it is to be used?

There are various ways in which validity is determined—that is, researchers have examined various *aspects of validity*. For example, some aspects of validity examine the extent to which a test represents the construct (an area to be measured) whose name appears in its title. Do the items in the test match the intent of the test? Another aspect of validity is used when a construct is quite complex. Consider the notion of "eighth-grade algebra proficiency," for example. What does it mean to know how to do algebra at this level? Let's think about this. There are a number of different attributes represented in algebra, including algebraic properties (such as the distributive principle of $a[b + c] = ab + ac$); order of operations (remember PEMDAS [parentheses, exponents, multiplication, division, addition, subtraction]?); laws of exponents; differences between real, imaginary, and complex numbers; trigonometric functions; and so on. All of these attributes are not necessarily of equal importance as far as eighth-grade algebra goes—that is, knowledge and ability to perform some of them are of greater importance than others. Thus, the validity concern refers to how the construct is defined and the representativeness of the sample of questions included in the instrument that is designed to measure it. Were more important categories represented in a proportional way that indicates their significance to the concept?

Standardized tests are also available for measuring attitudes, values, preferences, beliefs, and so on—although there are, perhaps, far

fewer of such measures. Typically, these tests are in a questionnaire format. They also have norms and show reliability as well as validity coefficients.

And now, another cautionary note. It is important to determine whether the test (or questionnaire items in the existing measure) is adequately aligned with your program and appropriately cover the concept intended by your particular evaluation question. The instrument may purport to have high validity, but still might not be *valid for your program*. You need to look carefully at the questions in the instrument—whether a test or a questionnaire. You might want to have the primary stakeholders sit with you to make that determination.

If only a partial coverage of the intended question is provided, perhaps the instrument could be supplemented by another measure. If, on the other hand, the instrument includes items that go beyond the question, and this is the best available measure, some evaluators suggest extracting information on just the test items available for the question. In essence, this is creating a new measure. If the measure had been standardized and reliability and validity calculated for the whole test, then they would not be applicable to your revised measure. This may not be of great concern if generalizing your findings to other places is not an issue, which it frequently is not. Moreover, it is not ethically appropriate to use existing measures that are for sale and to modify them. However, instruments that were used in research studies are typically available with the permission of the author.

Developing New Instruments

Let us now talk about the process for developing new instruments that yield quantitative data. I focus on the development of questionnaires and achievement tests and only briefly touch on the construction of other, less frequently used, measures.

RESPONSE FORMATS

I've commented earlier on the nature of a questionnaire. Let us now consider how they are developed. In essence, a questionnaire is an instrument consisting of a series of questions and other prompts that are designed to obtain responses typically related to attitudes, behaviors, points of view, perception, and so on. There are a variety of question formats used in questionnaires. I highlight two types: the *forced-choice format* and the *multiple-option format*.

The forced-choice format comes in three varieties: the *two-option variety*, the *multiple-choice variety*, and the *rating scale*. Responses to sur-

vey questions, regardless of which of these varieties are used, are considered forced choices because you can select only one response choice out of those presented. The two-option variety, for instance, is basically one that requires a "yes" or "no" answer, or a "true" or "false"—perhaps even "agree" or "disagree" (see Example K.1a). You are familiar, of course, with true/false tests. This option is appropriate when the question is factual and requires a dichotomous response (see Example K.1b).

Example K.1a. Two-Option Yes/No Questionnaire Item

Did you eat before going to work today?

- ☐ Yes
- ☐ No

Example K.1b. Two-Option True/False Questionnaire Item

I have lived in California.

- ☐ True
- ☐ False

The multiple-choice variety, on the other hand, provides an opportunity for the responder to select a *single answer from a number of possibilities* (see Example K.2a). This question format is often used to collect information about actions, behaviors, or other characteristics. When you develop the instrument, be sure to write the choices carefully so that they do not overlap. You do not want a respondent's choice to fit within two categories. This would be an issue if "pediatrician" or "triage nurse" were also included in Example K.2a, for instance, because a pediatrician is a physician who specializes in working with children and a triage nurse is a health care professional who works primarily in emergency rooms. Survey takers would be able to check off these options along with "physicians" and "nurses" if provided with these response choices. If you anticipate this being the case and want to force a single choice, provide "other" as a possible response and strongly encourage respondents to select only one response option (see Example K.2b). Otherwise, it is best to convert this question into one that uses the multiple-option format. You might even use this question format to determine knowledge, learning, or awareness. If so, it may be appropriate to also include "don't know," "not sure," "not applicable," or "decline to state" as a possible response choice (see Example K.2c).

Example K.2a. Multiple-Choice Format Questionnaire Item

Who do you most frequently turn to for health advice? (Select one option only.)

- ☐ Physicians
- ☐ Nurses
- ☐ Friends
- ☐ Parents
- ☐ Counselors

Example K.2b. Multiple-Choice Format Questionnaire Item with "Other" as a Response Option

Please indicate your gender identity. (Select one option only.)

- ☐ Female
- ☐ Male
- ☐ Other _____

Example K.2c. Multiple-Choice Format Questionnaire Item with "Don't Know" as a Response Option

The U.S. government funded research that led to the invention of the Internet.

- ☐ True
- ☐ False
- ☐ Don't know

Rating scales are also frequently used to collect quantitative data. Unlike yes/no or select-an-option formats, rating scales provide the possibility of gaining understanding about the degree or extent of an attitude or belief. It is not simply a question of do you agree or disagree, no or not no, believe or not believe, and so on. Rating scales introduce the notion of gradations. Typically, these scales are represented by 5 or 7 points. For example, in what is referred to as a *Likert scale,* respondents might be asked to indicate the amount of agreement or disagreement from "strongly agree" to "strongly disagree." Intermediate points would be "somewhat agree," "uncertain," and "somewhat disagree" (see Example K.3a). This kind of questionnaire format might also be presented more effectively as the rating option for multiple questions. Thus, the questions might be stated to one side of the page with a

5-point scale, which would be applied to all of the questions listed—as in Example K.3b. Note that while the examples highlight "level of agreement," the evaluator can also choose to gauge extent of adequacy, appropriateness, or frequency on such a scale as well.

Example K.3a. Measuring Level of Agreement with a Likert Scale on a Single Question

This book has contributed to my understanding of evaluation.

- ☐ Strongly agree
- ☐ Somewhat agree
- ☐ Neither agree nor disagree
- ☐ Somewhat disagree
- ☐ Strongly disagree

Example K.3b. Measuring Level of Agreement with a Likert Scale on Several Questions

Please indicate the extent to which you agree with the following statements.

This program has . . .	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
Helped me to improve my public-speaking skills.	☐	☐	☐	☐	☐
Provided me with the mentoring that I need.	☐	☐	☐	☐	☐
Been a valuable resource for my professional growth.	☐	☐	☐	☐	☐

There are a variety of other scaling techniques that carry with them different statistical analysis possibilities. Of these, the most familiar is the *semantic differential scale*. This scale is typically used for measuring attitudes or feelings. It consists of descriptive adjectives and their antonym at each end of the scale and seven "attitude positions" between them. Thus, one might have "good" at one end of the scale and "bad" at the other, or one might list "fair" and "unfair" at each end with seven

short lines between them that might be checked. Thus, in using this, you would measure degree of "goodness" (see Example K.4).

Example K.4. Measuring Adequacy with a Semantic Differential Scale on Several Questions

Please indicate the extent to which the following program resources adequately supported your growth.

	Adequately		Inadequately
Workbooks	☐	☐	☐
Handouts	☐	☐	☐
Tutors	☐	☐	☐

There are also comparative scaling techniques that provide the opportunity for items to be directly compared with each other. One such scale is called the *Guttman scale*. When using this scale, respondents are asked to indicate the extent to which they agree with a set of statements that are increasingly specific. Structuring items in this manner allows the evaluator to determine *at what point* the respondent no longer feels a certain way—a sensitivity analysis of sorts (see Example K.5). This scale and others of that ilk are more complex—look at some of these other possibilities if you wish. I would suggest, for now, that if you are using a rating scale, then the Likert format will usually suffice.

Example K.5. Measuring Agreement with the Guttman Scale

Please indicate whether you agree with the following statements.

	Yes	No
The activities offered in this program are helpful.	☐	☐
The tutoring offered in this program is helpful.	☐	☐
The Saturday tutoring sessions offered in this program are helpful.	☐	☐
The 2-hour Saturday tutoring sessions offered are helpful.	☐	☐
The 2-hour Saturday morning sessions offered are helpful.	☐	☐
The 2-hour Saturday afternoon sessions offered are helpful.	☐	☐

The second kind of question item format provides *multiple options,* and respondents can select as many as is applicable to them, or they can indicate all that apply (see Example K.6). This option is equivalent to allowing respondents the choice of saying "yes" or "no" to a number of questions in the format of a single item. All items checked mean "yes" or "agree" (or perhaps "no" or "disagree," depending on the wording of the question).

Example K.6. Multiple-Option Questionnaire Item

Which of the following individuals have been helpful to you in this program? (Check all that apply.)

- ☐ Program director
- ☐ Program coordinator
- ☐ Tutor A
- ☐ Tutor B
- ☐ Other _____

QUESTIONNAIRE CONSTRUCTION

Defining the question well is important. The evaluation question and what it means must be thoroughly defined in order to develop a questionnaire. We considered this issue in discussing the selection of existing instruments and their relevance to your evaluation. The same points must be made in discussing the development of a questionnaire. There are a number of guidelines that I suggest in questionnaire construction. First, *know your potential respondents* and their frame of reference. Where are they coming from? How might they interpret the question? Be sensitive to their feelings and how they might react to a question or possibly be offended by it.

The question should have a *clear meaning.* Ambiguous questions will yield meaningless data. Double negatives are confusing; avoid them. One serious mistake that many questionnaire developers make is inadvertently including more than one issue in a question—that is, writing a double-barreled item. For example, "Did you like the teachers and counselors in the program?" Does a "yes" response mean both were liked? Does a "no" response mean that neither was liked? What is the answer if you liked the teacher and disliked the counselor? In the same vein, stay away from leading questions. For instance, "Don't you think the teachers in this program are great?" These items bias respondents and are problematic because they generate inaccurate data.

The *writing style* and *presentation format* is also important. Don't get too academic. Is the wording simple enough to be understandable, or are there too many technical words? The writing style should be conversational. Furthermore, the way in which questions are formatted on paper should be appealing and not confusing. Don't cram questions on the page, because doing so creates a format that looks confusing or threatening. Be sure to leave white space on the page as well.

Question sequence is also important. Generally, think of questions as proceeding from more general to more specific, from most factual to most attitudinal and opinion related. Demographic and personal data are more sensitive, so place these questions at the end of the questionnaire.

In order for questionnaires to be reliable, they must be *administered* in a comparable way to all respondents. Clear, detailed instructions for administration are part of the questionnaire—an important part.

MEASURING ACHIEVEMENT

Perhaps you will find yourself in the position of needing to measure achievement but finding no achievement test that exactly fits the need of your evaluation question. You may be forced to construct your own *achievement test*. Such an instrument is a test of knowledge and developed skill. We typically think of achievement tests as related to information acquisition, the gaining of knowledge, and the understanding of facts or principles. Skill attainment more typically refers to abilities or proficiencies gained from training—as the skill related to riding a bike or knitting or learning a computer program, or whatever.

First, let me note that a number of things that I talked about in discussing questionnaires are helpful in constructing achievement tests. In many ways, questionnaires may be thought of as the generic type for collecting quantitative information. Achievement tests usually follow a questionnaire kind of format and might be considered a more specific type of questionnaire. Participants are asked to respond to a question, in which they provide an answer related to the content knowledge gained from participating in the program being evaluated. The format of those questions is similar to those that I talked about above in commenting on questionnaires—that is, we can determine whether those taking the test have acquired the necessary content knowledge by asking them true/false questions. For example, we can ask test takers to make a forced choice between two or more options that are presented to determine whether they have understood a particular concept, or we can use a multiple-option format in which respondents select all

of the options that apply to the situation. Thus, for example, we might ask those taking the test to indicate which of the choices presented are considered by health experts to be appropriate preventative procedures for a specific disease.

It is a difficult activity to personally develop an achievement test. When you attempt to develop a new measure of achievement, you forgo many of the advantages of a standardized, norm-referenced test. Clearly, you lose the ability to compare the results obtained by project participants with those attained by a norm group. This may be an important part of the way that standards for valuing are established—particularly in many summative evaluations (more on this in Section V).

As I previously noted, one of the major problems that you, as the evaluator, always face is whether the measure to be used fits the evaluation question. Are the objectives being measured those that you really want? Thus, there may be instances where you simply have no choice other than to develop an achievement instrument for use in your evaluation of a specific program. While you will not be able to expend the effort to standardize the instrument to the extent of developing norms, you will be better able to specify the attainments and the criteria that you wish to measure. The measure should ultimately better fit your program needs. There are more trade-offs than there are right or wrong answers.

Achievement Test Construction

In starting to develop an achievement measure, you will want to carefully describe each of the objectives to be measured. These should be so well specified that there is little room for ambiguity. In doing so, you will be able to develop specific test items that exactly conform to the objectives of your program. In contrast to norm-referenced tests, tests such as these are referred to by evaluators as *objectives-based tests* (or *criterion-referenced tests*). Since objectives-based tests exactly measure the objectives of *your* program, standards for judging may be tied to success on that measure. Thus, for example, one might establish as a standard that a stipulated percentage of participants had satisfactorily completed a specified percentage of questions correctly (e.g., at least 80% of participants achieved 80% or better on the test).

The development process of an objectives-based test begins with an understanding of the content standards—that is, what is the intent of the question related to achievement of learning within the program? What specific learning is implied within the question? This entails a great deal of understanding about what individuals are to learn or understand. You will recall that earlier, in developing the questions, I urged you to work with the primary stakeholders in fleshing out the

meaning of the question. If you did, you would have probed: "What constitutes an adequate answer? What are the components within the question?" Now, if you are developing a new measure to answer the question, these prior inquiries, while helpful, are probably insufficient. A great deal of specificity needs to be acquired when writing *clearly stated outcomes,* but understanding the desired outcome is even more complicated than that. Outcome indicators can be measured at different levels of attainment. One might gain understanding of a concept at a peripheral level—say, familiarity or general awareness—but what might be expected and required, and implied in the question, is at a far greater depth. Try to gain focus on the *level of performance* implied within the question. How deep an understanding is required?

As you seek to develop a listing of the subobjectives that constitute the question, you will soon become submerged in more questions than you are able to handle. I suggest that you work cooperatively with the primary stakeholders in considering *priorities* and weightings within those priorities. Then, focus on those subobjectives of greatest importance and relevance. Precise statements of the individual subobjectives of the program must be developed and, from that, test items would be constructed that measure the skills and knowledge deemed important. If there are more skills of greater importance, they should be represented proportionately by a greater number of test items. Typically, there are achievement attributes that are more easily attained by program participants. These should appear earlier in the test—especially if they are attributes that are needed to answer questions appearing later in the test.

A Word of Caution: I must repeat again an admonition previously made. You may not have the time and resources available to custom-make a test. Typically, such a task is beyond the resources available for many small- or medium-sized evaluations. However, if this is the only appropriate choice, proceed with care and caution.

RECAP—SECTION K

Quantitative Data Instruments

- Types of Instruments for Collecting Quantitative Data
  - Tests
  - Logs and checklists
  - Surveys

- Finding Existing Instruments
  - Tests and questionnaires
  - Standardized norm referenced
    - Reliability
    - Validity
- Developing New Instruments
  - Response formats
    - Forced-choice option (e.g., two option, multiple choice, rating scale)
    - Multiple option (e.g., select all that apply)
  - Questionnaire construction
    - Know your respondents
    - Questions have clear meaning
    - Writing style/presentation format
    - Question sequence
    - Administered comparably
- Measuring Achievement
- Test Construction (Criterion-Referenced Tests)
  - Clear specification of outcomes
  - Level of performance
  - Priorities and weighting

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

Keeping the RUPAS program and the evaluation questions that you developed in Section J in mind, let us now consider issues related to data collection. Review the various quantitative measures discussed in this section. When conducting the RUPAS evaluation, what instruments could you use? Have some of these instruments been developed by program staff, and are they already being used to collect data? Are these data available for your use as the evaluator? Who would be the best person to help you answer these questions—Amy Wilson, Carmen, Zoe, someone else? To what extent do *existing* instruments help you to answer the evaluation questions that were previously identified? If there are data gaps and you must develop new instruments, what might they include? How would these new instruments help to address the gaps in data? For example, if you needed to develop a survey, what might such an instrument be designed to shed light on? Or if you needed to design a test, what would it measure and among whom—students, parents, staff, or other stakeholders?

Resource

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.). (2017). *The twentieth mental measurements yearbook*. Lincoln, NE: Buros Center for Testing.

This yearbook series, compiled biannually, describes and reviews over 200 published tests.

Further Reading

Armstrong, P. (2016). Bloom's taxonomy. Retrieved from *https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy*.

Bloom's taxonomy of the cognitive domain provides an indication of the multiple levels of understanding applicable to the construction of quantitative measures.

Henry, G. T. (2005). Surveys. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 402–404). Thousand Oaks, CA: SAGE.

Questionnaires are one form of a survey. Gary Henry reviews among other things: survey design, developing and organizing questions, and administering such instruments.

Metfessel, N. S., & Michael, W. B. (1967). A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. *Educational and Psychological Measurement, 27,* 931–943.

This is an old article that focuses on educational situations, but it identifies all potential sources of data.

Quick Reads

1. Paul Bakker on Using Google Consumer Surveys for Evaluations
   http://tinyurl.com/zsmaesf
2. Shelly Engelman and Brandi Campbell on the Advantages of Using Retrospective Surveys
   http://tinyurl.com/zkg2sor
3. Lija Greenseid on Conducting Mixed-Mode Surveys
   http://tinyurl.com/zt47rf3
4. Rakesh Mohan and Amy Lorenzo on the Usefulness of Paper Surveys
   http://tinyurl.com/gtfwa5s

SECTION

L

What Are Instruments for Collecting Qualitative Data?

In the previous section, I talked about instruments that are typically used to collect quantifiable data. Now let us consider the ways that *qualitative data* are collected.

First, what do I mean by the word "qualitative"? You learned (or already knew) that the term "quantitative" indicates quantity (or numbers). So, do we therefore intuit that "qualitative" is some kind of boastful statement that implies that such data are of higher quality? Some qualitative researchers may think so, but that is not what the term means. Quality is meant to reference the attributes or nature of an entity. What are the qualities or *characteristics* of the entity? Qualitative researchers use their data to provide in-depth understanding of human behavior (individually or in programs); not necessarily describing what results were found or when, but rather, *why* and *how* it happened. Qualitative researchers believe that capturing stakeholders' experiences and perceptions leads to deep understanding of a program. Qualitative data help provide insights into a program by "carrying" potential readers into the program setting. The idea is that they can better understand and picture the richness and detail of what is happening in the program through the story that such data tell. There are a number of evaluator options that typically yield qualitative data. I now review some of these.

Before talking about specific instruments for collecting qualitative data, let me first talk about *data sources*. In Section K, I described

the different sources from which data could be acquired. Remember? I discussed the three choices of either acquiring existing data, using existing instruments, or developing new instruments. That theme is somewhat less relevant in this section. First, *existing instruments* is not a major topic when it relates to qualitative data. Surely, there are some instruments available, such as questionnaires, that might have open-ended questions, but these are primarily instruments that are used to gather quantitative data. As noted above, the strength of qualitative information is the opportunity provided for being especially responsive to the particular context and obtaining the perspectives of those within it. Thus, in this section, I do not comment further on the existing instruments category of data source.

Another category previously mentioned is *existing data*. There are a variety of types of data that might yield qualitative understandings. We discussed the various documents that might be available at the program site, or in the community, that yield quantitative data. Other artifacts containing text or narrative might yield similar understandings of the program as well. Think, for example, of the minutes of meetings, of photographs, or of work samples produced previously by program participants. E-mails, notes, or letters from program participants or program brochures are also among existing qualitatively oriented documents. Newspaper articles describing the program is another example.

What might you learn from some of these existing materials? How can such data be used? First, with "a grain of salt," so to speak. You did not originally collect these data. Rather, they were generated by others—the information contained in them has been "filtered" through the eyes of others. Therefore, you cannot attest to their accuracy or the context in which they were developed. Treat existing data, especially existing qualitative data, with care. These data provide possibilities and thoughts about issues that might warrant further examination.

Stop, for a moment. Before proceeding further, why don't you reflect on the kinds of documents, materials, or other artifacts that might be found in a program with which you are familiar?

DEVELOPING NEW INSTRUMENTS

For the most part, the primary data source for gathering qualitative data are new instruments. I next talk about the major types of instruments for gathering qualitative data—*observations* and *interviews*—including group (or so-called focus group) interviews. Each of these has specified procedures and guidelines for their accomplishment. However, there are options and variations in the way in which the procedures might

be implemented; I discuss this as well. Let me further suggest that to a large extent the instrument for collecting qualitative data is really the evaluator. Qualitative data involve your perceptions, your observations, and your interpretations; these must be carefully bounded. Appropriate training is essential for gathering qualitative data. This section provides a start in that training—but *only a start.*

OBSERVATIONS

Observations, also referred to as naturalistic observations, naturalistic inquiry, or participant observation, are methods used for getting more fine-grained information about a program, activities, participants, or other stakeholders. In terms of observations, does this really mean just watching other people? Well, yes, but it's more than that. Watching is a large part of what doing observation is about. The central idea of observations is to systematically engage in this careful watching until such time that continuing patterns or trends emerge. To start, you want to understand what is happening in as much detail as possible—as it is unfolding in front of you—as you are seeing it, hearing it, or maybe even smelling it. Then, you want to understand why. Why did what you see, hear, or smell happen in that particular way?

I believe that much of what you do and how you do it is structured by your role. The nature of the *observer's role* may vary from external (uninvolved) observer to participant. As a fully external evaluator, you do not have a role in the program and are present at activities only as an observer. In contrast, if you are involved in the program, you might have been enlisted (or volunteered) to engage in the evaluation as an evaluator-participant. In this role, you would be attempting to balance your participation by simultaneously observing and taking notes. Staff in programs and organizations with limited evaluation resources often find themselves taking on this role. Many evaluators, internal and external, strive to find a middle ground so that they can take on the role of a *participant-observer.* The role that evaluators play has implications for the detail of notes they might take, for their understanding of the context, their impartiality (or bias), or their better understanding of the implications of actions or events. What do I mean by this?

An evaluator-observer is more likely to notice patterns of activity and behaviors that might otherwise be overlooked by an evaluator-participant. In this case, evaluator-observers' lack of familiarity with the program helps them to be a bit more curious about what they are experiencing for the first time. While seemingly advantageous, their lack of familiarity also poses a challenge. The learning curve is

much steeper for the evaluator-observer than it is for the evaluator-participant. Greater effort has to be invested in the building of respect, credibility, and trust (remember this from Section F?). Relationships must also be nurtured from ground zero. All of these issues affect an evaluator-observer's ability to access privileged spaces and information, which influences the depth and quality of data that can be collected. The evaluator-participant, who is presumably very entrenched in the program, might be a bit more critical and reflective of what she or he is seeing. This familiarity, however, could also be a "blind spot" for evaluator-participants because it is challenging to question what might be considered common knowledge or common practice. There are advantages and disadvantages related to each position on the participant-observer spectrum. The message for you: Be aware of your position on that spectrum, the role that you have taken on (or has been ascribed to you), and the implications of that role.

I believe that the extent to which you, as the evaluator, can fully be a participant-observer varies according to the situation—that is, there will be times when you should, ideally, sit back and take in what is going on in the program. An example includes the beginning phases of the evaluation when, if you are an outsider, you are still trying to establish credibility, negotiate access, and learn about the program. However, as you gain familiarity with the program, the possibility and ease with which you can immerse yourself in programmatic activities becomes greater.

Observations vary in terms of *protocol style* and, accordingly, the quality and nature of the information that can be collected. What does this mean? Protocols are the system of rules and procedures for recording information that guide the way that you will go about your work. These protocols run the full gamut of being completely structured and inflexible (e.g., a checklist) to being completely unstructured (e.g., a blank notebook). Checklist-like observation protocols are used when evaluators are interested in documenting whether a prescribed activity was carried out or if they wish to tally the frequency with which a particular behavior was exhibited. Depending on how it was designed, this type of protocol leaves little room for taking notes and elaborating on what was observed. At the same time, it is advantageous because it allows for the efficient capture of data while in the field. The extreme case of this was discussed in the previous section where we considered observation protocols that are designed to yield quantitative data.

At the opposite extreme, observations may also take place in an unstructured fashion. This would allow you, as the evaluator, to capture what is striking to you at any given moment. This unstructured protocol provides opportunities to gain experiential data unfettered

by preconceived categories. Though the absence of structure allows for fluidity, it can also pose challenges because there is little guidance in determining what is noteworthy versus trivial. Instead, these decisions are left solely to the evaluator's personal judgment and may call for added effort if quantification of information is needed down the line. I personally prefer some middle position on this structure continuum.

Now let's talk about what you might record when you are observing. There are various approaches to *taking notes* while conducting observations. As noted above, you might use protocols or templates that guide how you will take notes—your decision will determine how you go about the note-taking task. If you use a checklist, you might be limited in how much you can write. However, what you get in return is the potential freedom to move around in the space, to observe from different locations, and to have different vantage points. If you opt for a plain notebook instead, you might be inclined to use the "write like mad" approach—firmly planted in one area of the room where you would be observing and documenting as much and as quickly as you can. Certainly, what you gain here is volume in the quantity of details that you can record and perhaps the amount of notes that you can generate and fill in later. What is sacrificed, however, is the level of mobility that is available to you. This prevents you from engaging with program participants and staff and precludes you from some of the activities in which you might have otherwise participated.

Alternatively, you, as the evaluator, may wish to exercise a subtler way of taking notes, which would involve jotting down key words or sketching images in a notepad that you might later use as *triggers* to write up fuller notes. This alternative approach, as you may expect, would allow you to be more flexible in terms of your movement and level of engagement in activities that are taking place around you. At the same time, know that you will be challenged to remember conversations, events, and other happenings that you observed—depending on when you choose to write up your notes. The cognitive burden that accompanies the need to remember large amounts of detail and to represent them accurately is real. Still, many evaluators prefer this approach to note taking because of the advantages that it offers over the others that I described.

So, how might you use these "triggers" to develop observation notes? Think about each of the key words or images and try to remember the context. What happened? How did it happen? Who was involved? Can you remember, more specifically, some of the specific words that were used?

Regardless of the note-taking method used, you must also be aware of several sources of bias that accompany observations. Bias are

those things that can influence your decision making in the field and how you understand or interpret what you experience. Your personal cultural context—including identity, values, upbringing, and professional training—all have the potential to inform and shape the evaluation work that you do. They might lead you to see only what you choose to see and only what is expected. Likewise, the program's context could also be sources of bias—for better and for worse. You might have a day in the field where you are not observing typical patterns of activity because a last-minute school assembly was scheduled, or because you are new to a classroom and curious students feel inclined to be on their best behavior to impress you. Alternatively, they might be more excited than usual by your presence and behave in ways that do not accurately reflect the typical energy in the room. In the same vein, program staff and instructors have personal cultures that might influence their actions. Like students and program participants, they might also unknowingly alter their behavior, thinking that it is what you—the evaluator—is expecting to see. These kinds of biases can cripple the quality of data being collected, but there are means of accounting for them.

The value of collecting observational data is the depth and richness of detail that you are able to capture. This cannot be accomplished in several hours of a day's time. Instead, prolonged and consistent exposure is necessary to gain a fuller understanding of what is happening and why, what is normal, and what is atypical. Prolonged fieldwork—even if only for 2 days—is what makes collecting observational data costly. Still, it is among the best ways to counter participant-generated bias. Furthermore, you might document your own biases and opinions of the activities as they are taking place by inserting your *observer comments* into your notes, or journaling about your experiences at the site, or writing memos about what might be causing bias. Explaining and accounting for issues that may color your perspective improves quality and adds to the validity of the data that you collect.

Thus, you would have two documents: the detailed notes that you recorded, and a separate set of observer comments. When these documents are put together, the whole set is called *field notes* and they become the source of information that you will use for further examination. Keep in mind, however, that you should not try to look for patterns as you are taking notes because you don't know what you are seeing yet. Just take notes. If you try to look for patterns right away, you will end up missing a large part of what's actually happening.

Observing is not easy. The development of *observer skills* requires training and preparation. Good observers should not only be astute.

Rather, they should also be familiar with the nature of human interaction in order to understand what is transpiring. The ability to concentrate fully on the activity being observed without distraction is vital. Also, good observers should have the ability to separate detail from trivia and describe what is seen vividly.

As with all data collection tools, there are advantages and disadvantages to engaging in observations. Observation data are a source of incredibly rich information that has the potential to help you fully grasp a context and the direct experiences of individuals, and to better understand their perspectives. On the other hand, this approach could be extremely time-consuming and costly. Moreover, it would require that you develop *your* observation skills, which takes practice. Thus, it is often helpful to complement data gathered using observations with data collected using other methods. Otherwise, it may quickly use up the majority of the evaluation budget.

INTERVIEWS AND FOCUS GROUPS

One of the ways that you, as the evaluator, might gain a deeper grasp of happenings within the program is to collect data by conducting interviews. Generally speaking, the goal of interview techniques is to capture the voices of different respondent groups. I previously have discussed observations and the skills required for using that instrument. These same skills are necessary in interviewing. The interviewer must be a good observer. When we talk to people, there are nonverbal cues. Are some responses given with greater ardor? Do some questions cause unease? What does that mean? Careful observation is an important part of interviewing.

Interviews can take place in a format ranging from fully structured to unstructured. In a *fully structured interview,* a preestablished protocol consisting of key questions is used. It guides the course of the conversation with little room for probing and adaptation. Structured interviews are designed so that they do not vary from one person to the next (or vary as little as possible). For this reason, interviewees are sometimes discouraged or even prohibited from asking questions of the interviewer. In many ways, this is like an orally administered survey but less efficient. Structured interviews may be appropriate when there is a need to have responses to the same questions from each interviewee. Situations that might require structured interviews are where there are a number of interviewees and there is a concern about the possibility of variability of response and unequal representation of issues addressed.

Semistructured interviews are also guided by a preestablished set of questions. However, questions do not need to be asked in a specified order—rather, interviewers are free to probe and shuffle questions as needed. This enables the process to be more conversational in tone. If there was anything the interviewee said that was unclear, the interviewer can ask for clarification and choose to go as deep into the question as she or he thinks is needed. Furthermore, both the interviewer and interviewee are free to revisit topics or issues previously discussed if further elaboration is needed. This process can be quite fluid compared with the structured interview.

What are the desired characteristics of questions to be asked in a semistructured interview? First, questions should be specific enough to focus on a single topic yet be open-ended enough to allow for multiple response possibilities. Questions should be written in nontechnical, neutral, and natural (conversational) language. It is best to field test or try out the questions before going to interview sites.

Appropriate sequencing is important. Recall that in Section K, I cautioned about not asking personal or sensitive questions at the beginning of a questionnaire. This is generally applicable in interviews, but is open to some variability because you want questions to flow naturally and sequence is partially determined by the way that natural lead-ins are provided by particular responses. It is important that you, the evaluator, are aware of the potential questions to be asked and not have to continually refer to your notes. In that regard, you should have thought beforehand about possible appropriate follow-up questions. Namely, the "why," "how," and "when" types of things. Semistructured interviews are often used when you have a good sense of the general topics that you would like to discuss and are more concerned with getting interviewees' perspectives than ensuring that questions were asked in a standard fashion.

Finally, we have *unstructured interviews,* or informal interviews. While there aren't any predetermined questions, the potential agenda is set by the interviewer. However, in this approach, respondents are allowed to say as much or as little as they wish. They are also encouraged to express themselves at their own pace and in their own terms. In that situation, you, as the evaluator, would sit down with someone—a potential source of information—and talk with him or her. But this "talking" can take on a number of different forms. The interviewee may bring up topics so that there's an element of "discovery," which lets things emerge during this kind of interviewing. In the more unstructured, open-ended type of interview, think of the therapist who asks a simple question and then mostly says, "I see," or "What do you mean by that?" and so on.

Focus groups may be considered as a special member of the interview family. The structure of focus group interviews may also vary as described above. However, focus groups do not occur in a one-on-one setting. Rather, they usually involve a group of between four and 10 respondents. Similar to interviews, a focus group moderator, possibly you, tends to have a set of questions to ask the group and the idea is that members of the focus group will have something to contribute or say in response to the question being asked. Once you feel that you have gotten a good amount of feedback or responses from the group, you might attempt to move the conversation on to the next question. But there really isn't any particular order in which questions are asked, because when you have a bunch of people sitting around talking, they'll typically initiate discussion on many of the issues that you had intended to address—you might even raise those that are not mentioned. As a whole, focus groups provide the benefit of enhanced understandings as a consequence of the interaction among participants. One person's comment provides the gist for another's and extends the idea.

However, disadvantages to this approach include uneven participation by those within the group; observing a phenomenon called "groupthink," whereby members of the group falsely come to share a common opinion regarding the issue at hand; and lack of control over the direction of the conversation. Taken together, all of these setbacks can translate into valuable time lost due to irrelevant discussion. Consider, as well, the issue of nonparticipation by those who do not feel comfortable making their views known in a group setting. This silence might be considered lost or missing perspective. Thus, while focus groups can produce incredibly powerful data, you will need to be a skilled facilitator to encourage participation when necessary, and to balance participation while limiting the degree to which conversations may head into uncharted waters. Furthermore, given the task of trying to control a group discussion, it is wise to conduct the group with two people—one person who facilitates the group while the other person takes notes.

SURVEYS AND QUESTIONNAIRES

We talked about *surveys* and *questionnaires* in Section K. They are usually thought of as tools that are designed to collect quantitatively oriented information. However, they can also be used to collect qualitative data. Features, such as *open-ended items,* can be built into these instruments that would be useful for qualitative data collection purposes. Open-ended items tend to be questions that respondents can answer in

a word, a phrase, or one to two sentences. Questions that may require lengthier answers are best suited for interviews and focus groups. When included on a survey or questionnaire, open-ended items serve several purposes. First, they enable respondents to provide feedback that is not bounded by numerical rating scales—they don't simply give a number, a ranking, or a rating. Second, by allowing respondents to comment on issues that they view to be important, they go beyond what is normally revealed in a quantitatively oriented survey.

As with all data collection methods—including open-ended items on instruments such as surveys and questionnaires—there are drawbacks. For example, respondents may not answer the question as phrased and use the space for different purposes altogether. They might provide added insights or voice their own gripes, but that too is useful. Answers given in this space may be revealing and provide further understanding and clarification. Responses may suggest trends that were not previously apparent and could lead to further investigation. Similarly, comments may support findings obtained from other sources of data. Typically, surveys and questionnaires are not used as the sole methods of collecting qualitative data. However, certain situations might force you, as the evaluator, to be dependent on them for qualitative data. Consider extreme budgetary or time constraints, for example. It is simply less costly to obtain a substantial amount of valuable qualitative information in this way. So whenever possible, I suggest that data collection tools be used in conjunction with other, complementary approaches. When doing so, it is also important to be intentional about the questions that are included on each. There should not be a great deal of overlap among interviews, focus groups, surveys, and questionnaires unless the questions that are asked on these are meant to aid in the linking of data across sources. It might be acceptable, for example, to ask for date of birth in several places if you need an anonymous identifier to connect a person's survey responses to her or his interview comments. However, if a respondent is already being asked to rate her or his program experiences on a scale of 1–10 on a questionnaire, then the same question ought not to be asked again in an interview. Rather, if possible, use the rating on the survey to probe about the quality of the interviewee's program experiences. Ask for examples that speak to the positive or negative elements that would have otherwise been too cumbersome to write about on the survey. The idea is to not only make data collection efficient and cost-effective but also to be mindful of the amount of work that you are asking your respondent to do. Those who participate in evaluative data collection activities are often volunteering their time so it is important to demonstrate sensitivity toward this.

RECAP—SECTION L

Qualitative Instruments

- Observations
  - Observer's role
  - Protocol style
  - Taking notes
  - Observer comments
  - Compiled field notes
  - Observer skills
- Interviews
  - Structured
  - Semistructured
  - Unstructured
  - Focus groups
- Surveys and Questionnaires
  - Open-ended items

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

Building on your responses to the RUPAS evaluation in the previous section, let's think about possible methods for gathering the necessary qualitative data (the recap might be a useful reference here). Specifically, would you collect observation, interview, and focus group data as well as program documents? How would you use each data type to answer the evaluation questions that you developed in Section J?

Okay, now that we have the ideal scenario in mind, let's introduce some elements of complexity since evaluation in the real world is everything but simple. Remember our initial budget conversation from the RUPAS evaluation in Section D (on contracting)? Let's now think about the amount of resources available and the ways in which each new type of data to be collected enhances our ability to answer the evaluation questions of interest—that is, what qualitative data might be collected if the evaluation was to be completed in 1 year and we had access to $50,000? How would things change if the budget was $150,000 spread over the course of a 2-year period?

Furthermore, think concretely about how the types and quantity of data that you will collect bears on the overall evaluation time line—that is, to what extent might interview data, for example, help you to answer stakeholders' questions with greater precision and accuracy? Might focus group data enable you to accomplish

this goal? How much of each sort of data do you need? Do you need both? What are the trade-offs (or pros and cons) involved in collecting the qualitative (and/or quantitative) data for your particular evaluation?

Further Reading

Krueger, R. (2005). Focus group. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 158–160). Thousand Oaks, CA: SAGE.

In this encyclopedia entry, Richard Krueger—one of the leading scholars on focus groups—presents a concise description of that method.

Mabry, L. (2003). In living color: Qualitative methods in educational evaluation. In T. Kellaghan & D. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 167–185). Norwell, MA: Kluwer Academic.

This is an excellent, easy-to-read discussion of the qualitative data collection and analysis process.

Patton, M. Q. (2003). Qualitative evaluation checklist. *Evaluation Checklists Program*. Available at *irantvto.ir/uploads/qec.pdf*.

This is an excellent checklist of things to think about in qualitative evaluation.

Ryan, K. E., Gandha, T., Culbertson, M. J., & Carlson, C. (2014). Focus group evidence: Implications for design and analysis. *American Journal of Evaluation, 35*(3), 328–345.

This article offers guidance on how to design and implement focus groups.

Quick Reads

1. Moya Alfonso on Becoming a Well-Trained Focus Group Moderator
   http://tinyurl.com/j869uuy
2. Susan Kistler on Formatting Qualitative Questions for Online Surveys
   http://tinyurl.com/z7vxr6e
3. Michael Quinn Patton on Purposeful Qualitative Sampling
   http://tinyurl.com/j9pfeqr
4. Janet Usinger on Interviewing People Who Are Challenging
   http://tinyurl.com/hzg2hxy

SECTION

M

How Do Data Collection Issues Impact Potential Evaluability?

What are "data"? (Note the seemingly peculiar use of a plural "are" verb with the word "data." In fact, "data" is the plural of "datum." Thus, as awkward as it may sound, we use the plural.) Now, to continue, data are information. They are information that have been collected by an evaluator in a systematic fashion. So where do these data come from? How does an evaluator know which data are most relevant?

First, I ask you to consider the notion that evaluators are surrounded by data. Indeed, almost anything could be considered data. Think of a substance abuse program. What kinds of data might be required in an evaluation? Do people cease using drugs permanently or for some specified period of time? Do people attend required sessions? What are their attitudes toward drugs? Do they physically feel better? Do those who know them feel that improvement has taken place? What are their feelings about the content of the program? Are program participants interacting positively with one another? Is the program really a viable one? Are the program activities those that were intended? There are more evaluation questions that might be asked. The world of a program is filled with data and with questions. One issue is *which data are relevant*?

AGAIN, BE CLEAR ON THE QUESTIONS

In order to know which types of data are relevant, evaluators must be clear on the evaluation questions that they are being asked to

answer and why individuals with a vested interest in the program—stakeholders—are asking those specific questions. Therefore, not only are purposes for evaluating and evaluation questions highly interdependent, but evaluators are responsible and obligated to assist stakeholders in articulating, specifying, and prioritizing the questions the evaluation should address. (I know that I have been harping a lot about "questions"—it's really important.)

Evaluation issues to be addressed, as seen from the list of possible questions noted above, can be quite diverse. The important thing in defining data requirements is to identify the focus of the data to be acquired. What evaluation questions are being asked and what are the data sources related to such questions?

FOCUS OF THE DATA

So how should you approach this? First, it is important to consider the potential focus of the data—let's examine the various possibilities. The issue: *Who or what is the focus of the evaluation question*?

Questions of concern, for example, might focus on the program participants. Let us consider a school situation. We might want to know about the students. Did they perform well academically? We might also wonder whether the program is having an impact on program staff—teachers, for example. What about the impact on other people in or out of the program? Parents or the community at large may fall into this category. Finally, the focus of the evaluation and the area in which data are to be acquired might be the program itself or even discrete activities within the program. Are these activities functioning? Are they having the desired impact?

Data might be acquired *directly* or *indirectly* about each "who" or "what." Now what do I mean by this? Directly refers to obtaining the data *from* those who are the focus of data acquisition, whereas, indirectly refers to obtaining data *about* them—that is, if the focus is students, then direct acquisition implies that you, as the evaluator, might obtain the data from the students themselves. On the other hand, indirect acquisition implies that you might go through another source for such data.

PROGRAM PARTICIPANTS

Let us unpack this a bit further by considering ways in which data may be gathered directly *from* program participants. Direct contact occurs in

a variety of fashions. You, as the evaluator, could *interview* students (if you will permit me to continue with the same example). Or you might collect information directly from these student participants through the use of *questionnaires*. Data might also be acquired directly from participants by *self-report*. In this instance, the evaluator does not personally ask questions of participants. Rather, the evaluator might rely on logs or diaries that students maintain about their program experiences. The third form of data acquisition from participants themselves is *performance data*. This is the form with which we are most familiar. The classic example of this data form is tests given at the end of a program to determine the extent of specific skills acquisition. Think, in a school example, of reading test results.

Now what about collecting data indirectly *about* program participants (students, in the example we have been using)? There are several possibilities here. For example, we might ask those who have ample opportunity to observe them about the extent to which they are performing appropriately and well. In the school example, we might contact staff and interview them about the students. We might ask teachers to respond to a questionnaire about the students, or we might examine teacher notes or comments about students, but it is not only teachers and other program staff who are in a position to provide data about students. Think also about other stakeholders, including administrators, parents, and community members. For broader data considerations—about the program itself and its operation—program funders and others might be thought of as additional stakeholders and be a valuable source of information about program participants.

Documents are another valuable source of evaluation data about the participants. Some preexisting documents that are available were created for a prior particular purpose. Such data that are extracted from archived documents are information that is revisited to determine whether they are relevant for current purposes. Examples of these types of data typically consist of various records stored in administrative databases—such as attendance records or prior-year standardized test scores. Documents might also include collections of photographs, recordings of program events, and so on.

Another source of data about participants (also indirect) is *observation* of those who are participating in the program being evaluated. In this instance, we would not ask participants to comment about themselves—rather, we aim to gain knowledge through passive observation. The evaluator is observing behavior to determine whether, for example, knowledge has been acquired, attitudes changed, or interpersonal relationships modified.

PROGRAM STAFF

Let us also consider "staff" as the focus of the data. Program staff could, of course, provide data directly about themselves. This could be provided through an interview. One topic might be the staff competencies that were presumed to be developed during the course of the program. Data on staff might also be acquired indirectly. Participants might be asked to rate staff performance. Program documents in the form of various administrative records might be examined. Perhaps you, as the evaluator, might want to gain information by observing the staff in action, as they perform their duties.

THE PROGRAM

I do not want to unduly prolong this discussion; however, since programs are the foci of many evaluations, a brief comment might be appropriate. Let me illustrate what I mean and how related data might be acquired. The question might be whether the program seems to be working—that is, functioning well. To obtain such data, we might seek the impressions of participants, staff, or other stakeholders. Program documents might provide insights into this question. Such documents could include reports on the program's receipt of materials and implementation of program activities or progress of participants generally. Finally, we might systematically observe the operation of the program.

SELECTING INDIVIDUALS

I have discussed with you the sources from which data might be acquired. However, when one of the sources is *individuals* (e.g., participants, staff, or other stakeholders) another question arises. Do we collect information from *all* participants? Is it possible, for example, to interview all stakeholders? How do we select which individuals will constitute the source of data? Clearly, when only a subsample of a total population (e.g., all of the staff) are to be interviewed, there is a possibility that the data collected might not represent all members of the group. Information might be biased in one direction or another. What can we do?

Sometimes potentially quantifiable data, such as data from tests or questionnaires, will be collected. In these kinds of instances, a wide range of categories of sampling techniques have been described by researchers. One such technique is referred to as *random sampling*. This

means that everyone in the total population has an equal chance of being included in the sample. In essence, the technique requires that the people to receive the questionnaire should be chosen in some random fashion. This might involve numbering all participants and then using what is called a random number table. Statisticians have developed procedures for determining the minimum number of people to be sampled in order to affirm that the data collected from the random sample truly represent the total sample.

We conduct sampling to enable us to answer evaluation questions. Thus, the nature of the question should help to determine the kind of sampling that will be done. If, for example, you were concerned with seeing the differences that the program makes on men versus women, you might want to use a *stratified random sampling* approach to ensure that an equal number of men and women are selected—that is, first develop a list of men in the program and women in the program, and then randomly sample each.

In another situation, suppose that you wanted to know, specifically, how low-income students were doing in a school program. In that case, the sampling should adhere to that focus—low-income students. Then the data to be collected might be *purposive,* and you would choose to sample from that particular population.

A potential confounding effect related to your gathering of what we now call quantitative data is the reality of who actually participates. Were those selected to be sampled, in fact, representative of whom the data were collected from? Thus, if a survey or an achievement test was administered within a program, were all respondents, or all intended respondents, present? Was there a systematic reason for those who were not present? Were, for example, participants who tended to do poorly not present? If a questionnaire had been sent to a prescribed population, it is important to know the percentage of people who returned the survey. Researchers and evaluators refer to this as the *response rate*. It is important that there be a sufficient response such that we can have confidence that the data elicited are really representative. Statisticians differ about what is considered an adequate response rate. Depending on the number of people to whom the survey was administered—on the order of several hundred, for instance—certainly one should expect at the very minimum a 50% response rate, but 60–70% would be the preferred minimum. If the survey was administered to a program of 10–12 participants, on the other hand, as close to a 100% response rate would be ideal. Again, statistics books can be consulted for more detail on an appropriate response rate for different instances.

Sampling related to the collection of data by nonquantitative methods is more difficult to describe. Types of sampling have been variously

described by evaluators, researchers, and methodologists: deviant case sampling, maximum variation sampling, homogeneous samples, typical case sampling, critical case sampling, snowball sampling, and so on. It really all boils down to "What is the question?" A concern for examining major successes or major failures may lead to using *deviant case sampling*—picking individuals at the extremes of the population. The desire for a great deal of heterogeneity in order to capture major themes may lead to choosing a very diverse group—*maximum variation sampling*. A concern for understanding a particular subgroup in depth may direct us to select *homogeneous samples*. When the most commonly found (or typical) individuals or group is selected, that is *typical sampling*. Sometimes because of the political situation, there is a need to select a particular case of greatest interest—*critical case sampling*. Finally, in interviewing, you might want to use a technique of gaining referrals from interviewees of those who they believe would have the most information and be most helpful. This is referred to as *snowball sampling*.

And so, in considering data to be acquired from interviews, observations, or focus groups, one needs to be guided by the question under consideration and, thus, there are many ways in which samples might be selected.

GAINING DATA ACCESS

I have discussed what data are and where data can come from to set the stage for examining issues that may arise in the process of data collection. You want to be sure that there will not be hang-ups—actually, there always are, so you want to minimize them. Consider for a moment the potential problems. What if you have difficulty gaining access to those from whom you need data—be it interviews or administering a questionnaire? What if there are scheduling problems? How do you work through problems around poor-quality data? How can you satisfy the organization's concerns about protecting the privacy of participants? Questions, questions, questions. Let's talk about this.

Gaining access to data is, unfortunately, a common problem for many evaluators, and is probably most often found in large bureaucratic organizations or institutions. In these types of settings, it may seem to evaluators that they are constantly working through red tape and negotiating, negotiating, negotiating. Indeed, data access is one of the more frustrating aspects of evaluation even with careful attention to prior discussion (and assumed agreement) about issues of access and rules for negotiating disputes.

So how might you, as the evaluator, go about negotiating access in a practical way? What are the access steps—specifically? What might be the potential barriers related to each step? The first step involves speaking to primary users and making both sides' *expectations* around data access *explicit*. Doing so successfully means that evaluators and stakeholders must understand which particular data points are needed, for what purposes, and when. This should be done in the early phases of the evaluation. For example, when evaluating educational programs, it is common for evaluators to need data about students' academic performance (e.g., course grades, exam scores) and demographic information (e.g., grade levels, gender, and ethnic identity). However, you might receive "pushback" from primary users because the importance of certain kinds of data in answering questions may not yet be apparent to them. Addressing this point of tension calls for the need to help stakeholders understand how these data will be used and stored. Certainly, it would be wise to reemphasize that data will be used only to answer the evaluation questions (no more, no less) and that the data will be stored in a secure location, to be accessed only by relevant members of the evaluation team. It would also be helpful to clarify how often you will need to receive this information—whether it is once or twice per year or more often, these expectations should be made clear early on.

While we have discussed examples of data to which access is important, it is equally critical to explain what data points you *do not need*. Social Security numbers and addresses are a few common examples. You see, one of the things that organizations are concerned with is the protection of privacy (we elaborate on this later in this section), so demonstrating that you are aware of their priorities is helpful. This is also an opportunity to set the stage for mutual learning by both sides—that is, what is needed to support the evaluation's success and genuine limitations. Nonetheless, the ground rules for access should be addressed and agreed upon. It may also help to consider writing them into the evaluation contract as is discussed in Section D.

The second step involves *establishing and maintaining relationships* with those in the organization who are considered information *gatekeepers*. You may accomplish this by assessing gatekeepers' potential roles and interests in the program and the evaluation. Then, maintaining open lines of communication is important because doing so helps to ensure gatekeepers' understanding of the purpose and scope of the evaluation. Fostering collaborative, mutually respectful relationships is also important. In this way, you and they are more likely to be on the same page, the same side, and "speak the same language." Sometimes it may be challenging to communicate with gatekeepers. Perhaps they are time challenged and view you as a further intrusion. They simply

may not want to work with you or may possibly view you as a threat. Try using different modes of contact. E-mails are used widely now, but picking up the phone is just as easy. Don't overburden the relationship; unnecessary communication can turn off busy people. Or, if necessary, consider gently tapping the organizational hierarchy and raising your concern in a "non-finger-pointing" way. However, remember that you, as the evaluator, will need to be judicious when doing this because you run the risk of jeopardizing potential future relationships with gatekeepers and other program staff.

COLLECTING DATA

Let us talk first about the things that you will have to do in order to collect data in the field. To collect both quantitative and qualitative data, you will need instruments. We talked about the different kinds of instruments in Sections K and L, but consider for a moment the *administrative preparations* needed to collect data. In the case of questionnaires, for example, consider whether you will have to administer them yourself or whether program staff will be assisting you. If you will be receiving help, who will be helping you? When and where will the questionnaires be administered? Who do you need to coordinate with to make these arrangements?

As for observations, it again goes back to the issue of negotiation. This applies to all kinds of sites, from the typical classroom to offices to other places of business (even public places such as parks!). From whom should you seek administrative permission? Is there one "right" person to ask? The important thing here is to realize that when in the field, the "right" people who can grant you access may not be obvious and that there may be different levels of access. For instance, if you are doing observations in a school cafeteria, while you certainly should clear this with the program manager, might it not be appropriate to also ask the cafeteria manager for permission to observe? After all, as an evaluator, you are entering someone else's workplace and should demonstrate respect toward those who live out a part of their lives in that place.

What does seeking *permission* directly *from participants* involve? Well, you should tell participants who you are, what you are trying to accomplish, how you intend to go about your activities, and ask if they would be willing to help you through the process. All the while, be mindful of the language that you use and how you interact with participants. As in reporting, avoid jargon and overly technical language—not only because participants might not understand but because you

don't want to be misunderstood as being condescending. At the end of the day, realize that some participants may not allow you access. While that is disappointing, it is more appropriate to respect their wishes and exclude them from your data collection efforts. Just remember that when you are working with people, one of your primary obligations is to do no harm (yes, as in "first, do no harm" from the Hippocratic oath).

In interviews—particularly focus groups—there are *space considerations*. Think about where the focus groups will be held. How many rooms will you need? Through whom do you reserve the space? How will participants know where to go? Do you need signs? Will you be coordinating directly with participants, or, again, will program staff be assisting you in any way? In the same vein, will the rooms be conducive to audio recording if doing so is necessary? The acoustics in some spaces will result in echoing, whereas others do not. What kinds of rooms will you be in? These are the kinds of logistical issues that need to be considered when arrangements are being made in preparation for data collection.

Much can go wrong in attempting to collect data. Despite all your attempts at careful planning, things happen. A number of things could affect scheduling: weather, attendance, an unusual program disruption. Evaluation work may need to be rescheduled. Try to *anticipate possible impediments* and consider ahead of time some of the ways that you might respond if the "what if's" happen.

QUALITY OF DATA

Let's suppose that you have been able to collect the necessary data. What will you do when you realize that the response rate for the last questionnaire administered was only 20%? On the other hand, what if you were able to collect all of the questionnaires only to notice that there was a lot of missing data because participants did not answer most of the questions? Here we have the same problem in that you have data that can't be analyzed, and typically, if you can't analyze the information you collected, then you can't use it. Game over.

Again, you can potentially avoid many of these problems if you *anticipate possible impediments* and possible areas for improvement before you go out to the field. Consider, for example, the instrument that you use, participants' characteristics, and the timing of your data collection. In terms of the instruments, assess their content, language, and length. Were the items tapping into relevant issues? Was the language used free of jargon and accessible to your audiences? If you are using context-specific language, are you doing so in a manner that is

aligned with participants' use? Is the instrument of reasonable length for the participants? Remember, the more complicated and lengthier the instrument, the more time it'll take participants to complete. Respect their intelligence and their time.

Next, determine the characteristics and *needs of participants*. Specifically, think about whether your instrument needs to be translated. Do your instruments ask about topics that are taboo in other cultures? Are the items highly controversial? Do the questions unduly lead respondents to a particular response? Pay heed to the degree of clarity, cultural sensitivity, and possibility of biased responses in your questions.

It is likewise important to be mindful of who you are and how others might experience you as the evaluator. Consider, for instance, your own experiences and cultural lens and how they might affect your data collection efforts (more on this in Sections K, L, S, and T). Evaluators who work across cultural, economic, and political boundaries might consider collaborating with program staff or other colleagues in this process, particularly if there is an interest in accounting for assumptions or biases that could potentially affect the evaluation's quality and accuracy.

Finally, consider the *timing* of your data collection. When working with schools, for example, think about whether data collection activities are unreasonably disruptive. Are they taking instructional time away from students and teachers? Be aware of the school's regular schedule and of any special events that might impact data collection. How can you incorporate data collection into the day-to-day routines of the classroom so that it is woven in seamlessly? In places of business with atypical hours of operation or untraditional organizational structures, consider adapting your activities to the organization's schedule and priorities.

UNDERSTANDING THE ORGANIZATION'S VIEWPOINTS

Sometimes, evaluators' expectations do not coincide with the organization's perceptions of what is best. We mentioned earlier that organizations feel that they are obligated to *protect the privacy* and *confidentiality* of the people whom they serve. So even though evaluators' services are sought out, it is misleading to think that evaluators are necessarily entitled to full, open access to sensitive information. At the same time, the program you are evaluating is obligated to not make it unreasonably challenging for evaluators. Rather, evaluators and organizations should have measures in place that outline the circumstances under

which access can be granted and what the data's explicit uses can be. That's one reason why organizations interested in regulating research and data use create offices or procedures for evaluating the acceptability of research and evaluation; these are typically called *institutional review boards* (IRBs).

An IRB is a committee that consists of a diverse group of researchers and community members. The committee is responsible for reviewing, approving, and monitoring proposed research studies that require involvement of human participants. Specifically, such a board is interested in protecting the rights and well-being of these participants. Most studies submitted for IRB review are from the medical, biobehavioral, or social science fields. While evaluation studies differ from research studies, they may still be subject to IRB review, given the often extensive role that program participants play during the evaluation process.

In addition to protecting the rights and confidentiality of participants, those in charge may have *program interests* that run counter to your possible needs as an evaluator. Organizations are focused on what they can do for program participants. Evaluation is often viewed as an unwelcome intrusion. In their view, time spent on evaluation is time that could be used for program activities. Also, program staff may misinterpret evaluation of the program as an evaluation of their performance, or they may think the evaluation is being conducted to make decisions about funding or for reorganizing the program structurally.

In all cases, evaluation may seem threatening and it is the evaluator's job to assuage misunderstandings and reduce threat. This can be done by appropriately engaging participants in the evaluation process, and by explaining to participants the purpose and scope of the evaluation along with the role of evaluation activities in which they are engaged.

Finally, always try to *balance* your *needs and interests* with those of the organization and program stakeholders. Try to minimize the potential burden that the activities you want to carry out will be for participants. Remember that even though your job is to evaluate the program, you do not want to cause harm or disruption.

➢ ***My Advice:*** Try to anticipate all that might go wrong and be obsessive in attending to the details of data collection. (You probably have already guessed that I am pretty obsessive, by nature.) Some further rules of thumb: build a close, respectful relationship with primary stakeholders; be prepared for changed circumstances; and negotiate the best possible solutions.

It is also wise to have some familiarity with federal privacy regulations, such as the Family Educational Rights and Privacy Act's (FERPA) and the Health Insurance Portability and Accountability Act's (HIPAA) rules. We are not suggesting that details of these regulations be committed to memory. However, many organizations are bound by these laws and awareness of them may be helpful during discussions about data access and data collection.

RECAP—SECTION M

Data Collection and Evaluability Issues

- Be Clear on the Questions
- Focus of the Data
  - Who or what is the focus?
  - From where are data acquired?
  - Gathered directly or indirectly?
- Types of Sampling
- Gaining Data Access
  - Make expectations explicit
  - Establish relationship with gatekeepers
- Collecting Data
  - Make administrative preparations
  - Seek permissions from participants
  - Consider space issues
  - Anticipate possible impediments
- Improving Quality of Data
  - Anticipate impediments
  - Be aware of needs of participants
  - Consider appropriate timing of collection
- Understanding the Organization's Viewpoints
  - Protecting privacy and confidentiality
  - Submit to Institutional Review Board
  - Recognize the priority of the program's interests
  - Reduce threat
  - Balance needs and interests

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

Our discussions about the RUPAS evaluation thus far has brought us to the point where it is necessary to consider issues of logistics and evaluability. So, let's begin by thinking about some potential difficulties that you might face related to the people. Who are the program participants in the RUPAS case? In considering the evaluation questions that you gathered from the RUPAS program stakeholders, are "participants" defined as the children, parent leaders, staff, or all of these? From whom might data be acquired?

Next, let's consider the location of program sites. Do we know the distance between various sites? How would distance influence how you go about data collection? In the same vein, if you are an evaluator who is not based in the Pacific Northwest, how would you manage communications with RUPAS program staff and with others who would participate in the evaluation? If you are indeed a local evaluator, how might your approach differ?

Are there other possible data collection–related challenges? What about those concerning language and culture? Do you speak Spanish—the dominant language used within the migrant community? If not, how would you overcome this obstacle? For instance, would you assemble an evaluation team with a Spanish speaker on board? Or would you hire an interpreter?

It is possible to come up with many other scenarios in which the evaluation planning can go awry. Can you think of other examples and potential contingency plans?

Further Reading

Donaldson, S. L., Gooler, L. E., & Scriven, M. (2002). Strategies for managing evaluation anxiety: Toward a psychology of program evaluation. *American Journal of Evaluation, 23*(3), 261–273.

This article provides a discussion of the impact of stakeholders' evaluation anxiety and presents some strategies that help to overcome this anxiety. Clearly, this is related to less traumatic data collection.

Fitzpatrick, J. L. (2005). Human subjects protection. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 188–190). Thousand Oaks, CA: SAGE.

Jody Fitzpatrick does an excellent job of describing the extent to which evaluators need to act in ways that respect the rights of participants.

Hatry, H. P., & Newcomer, K. E. (2015). Pitfalls in evaluation. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of practical program evaluation* (pp. 701–724). San Francisco: Jossey-Bass.

This chapter presents 27 potential pitfalls in evaluation, divided into before, during, and after data collection.

Jacobson, M. R., Azzam, T., & Baez, J. G. (2013). The nature and frequency of inclusion of people with disabilities in program evaluation. *American Journal of Evaluation, 34*(1), 23–44.

This article describes how, when, and the nature of participation of individuals with disabilities in the evaluation process.

Taut, S., & Alkin, M. (2003). Program staff perceptions of barriers to evaluation implementation. *American Journal of Evaluation, 24*(2), 213–226.

This article seeks to provide an understanding of the logistical difficulties of data collection—and evaluation implementation generally—by examining program staff perceptions.

Quick Reads

1. Linda Cabral and Judy Savageau on Improve Your Surveys by Conducting Cognitive Interviews
   http://tinyurl.com/gms8jaw

2. Lindsey Dunn and Lauren Fluegge on Organizing Data
   http://tinyurl.com/he96g63

3. Lisa R. Holliday on Using Data Dictionaries to Improve Data Quality
   http://tinyurl.com/hkxdkej

4. Dan Jorgensen on Effective Data Management
   http://tinyurl.com/zcmfa6k

SECTION

N

Are the Questions Evaluable?

The issue here is whether the questions can be answered. Are they *relevant*? Are the answers needed? Let me review where we are at this point.

First, you have gained a better understanding of the program and its logic. Then, some potential evaluation questions were specified. In doing so, you did an *initial consideration* of whether the program can be evaluated and answer the questions that stakeholders expressed interest in. Furthermore, you examined potential data sources, both quantitative and qualitative, that might be appropriate for answering the stipulated questions, and, hopefully, you now have some knowledge of the political, social, and organizational context surrounding the program. This will be helpful in determining the evaluation's feasibility. You, as the evaluator, have several tasks in determining whether the potential evaluation questions can, in fact, be answered. Now, we turn to these issues.

There are a number of reasons why the evaluation questions may not be evaluable (able to be evaluated, or able to be included as questions in the evaluation). Let me break this into seven categories: (1) *nature and stage of the program,* (2) *availability of resources,* (3) *nature and relevance of the question,* (4) *standards for judgment,* (5) *technical issues,* (6) *ethical concerns,* and (7) *political feasibility.*

STAGE OF THE PROGRAM

Do the nature and stage of the program offer insights into potential evaluability? Is it premature to ask this particular question? We need to be sure that the program is at a sufficient stage of development so that the answer to the question can be ascertained. If, for example, the program is intended to be 3 years in duration with particular outcomes expected at the end of that period, then it would be inappropriate (or should I say, foolhardy?) to attempt to measure final outcomes before the 3 years have expired. There must be sufficient time for the program to have operated so that the effects can be accurately assessed. Still, many stakeholders will feel pressed to gain insight into their program's quality in one way or another. Your role as the evaluator in such a situation would be to facilitate the identification of appropriate and evaluable processes and outcomes *in light of* the program's history and stage of development. Do not set the evaluation—and yourself—up to fail by attempting to find evidence of long-term success and impact if conditions do not allow you to do so.

RESOURCES

The next consideration is resources. What resources are available? This question will need to be answered in greater detail in developing the evaluation plan, but some initial estimates are certainly called for at this stage. Resources are of two types: fiscal and "in-kind." As in much of life, aspirations frequently exceed resources. We want more than we are able to afford. So, begin by checking carefully about the number of dollars available for the evaluation. Consider what it costs to conduct the evaluation—staff, time, materials, travel, and so on. Can an evaluation based on the selected evaluation questions be done? Especially, can it be completed in an acceptable manner?

Resources not only include the dollars available for funding the evaluation—rather, they include "in-kind" resources as well, such as availability of program staff time for certain kinds of assistance (e.g., data collection or clerical help). There is also the matter of equipment or other services to be provided by the client's organization (e.g., printing, computer access, work space). Determine the extent to which the evaluation will receive needed cooperation and assistance from the client and other stakeholders. How will they cooperate? What program resources will be made available?

Answering some evaluation questions will require a great deal of resources. We discussed the affordances and shortcomings of various

data collection tools and how they can affect the evaluation in Sections K–M. Those issues resurface here, in the context of evaluability. We already know that some data, and hence data collection approaches, are better at elucidating certain issues than others. Well-designed surveys, for example, are better and more cost-effective at shedding light on the *typical* program experience, whereas interviews grant insight into *particular* experiences, and they do so at a premium. Just as there are trade-offs between different data collection methods, sacrifices concerning evaluation questions can be expected as well. Some questions may require the consumption of a substantial part of the budget to the detriment of being able to respond to other questions. Depending on the resources available, allocating personnel, financial, and material support to answer evaluation questions about outcomes and long-term impact could mean that information about processes and short-term gains are de-emphasized or completely lost. These trade-offs must be seriously weighed against one another. Is the gain at the expense of another worth it?

NATURE OF THE QUESTION

First, there is the evaluation question itself. We have dealt with this in Section M. Ideally, at this stage, the question had been sufficiently refined so that it is a real question, rendering the two concerns I state below as irrelevant. Review again the evaluation questions and issues, and make sure that you have a firm understanding of what they are. If you do not clearly understand the question, then take the necessary steps to gain that understanding.

The first concern is whether the *question* that is to be potentially investigated is of insufficient worth. On the one hand, it might be so trivial that an evaluation based on that question is hardly worth the effort. Alternatively, the answer might already be known, or, if unknown, is the answer so simply acquired that it need not be a focus of the evaluation?

Second, we need to again ask ourselves (and stakeholders), "Is this question useful?" Is it a question that someone *wants* an answer to, or is it one that someone *needs* an answer to? Indeed, wanting something and needing something are quite different. We must understand whether the question is useful and if the answer is wanted because someone intends to use the information for program improvement, or to make a judgment about the program—that is, how will the answer be used? Is this question worth your time as an evaluator, or is it irrelevant? I return to this issue of want, need, and intended use in Section

T (on how analyzed data answers questions) and Section V (on how evaluators help evaluations to be used).

STANDARDS FOR JUDGMENT

I like to consider the possibility of establishing standards as part of the process of considering evaluability. Recall that in Section I, I first discussed the process of working with stakeholders to determine questions of real interest. In part, this is done by creating scenarios describing possible results for each question, and asking stakeholders about the extent to which those hypothesized findings would be useful. Description of further gradations of those possible findings served to determine the "value" of potential evaluation results. In essence, this process, when followed through, helps to provide standards for judging the outcomes of the evaluation—that is, it helps stakeholders to determine what is "good," "bad," and "good enough." Authors of one of the further readings to this section refer to this as "front-end-loaded standard setting." Similar to other aspects of evaluation, it is reasonable to anticipate that the standard setting process will require a fair degree of negotiation among stakeholders and between stakeholders and the evaluator. What comes to bear in these discussions is the reconciliation between stakeholders' historical understanding of the program and the evaluator's prior evaluation experiences. You will note, in Section T, that there are myriad ways that standards can be set and for data to be ultimately valued, but I like the idea of stakeholders participating in this manner to test the importance of evaluation questions.

TECHNICAL ISSUES

Of course what readily comes to mind in considering whether a potential evaluation question is appropriate for inclusion in the study are the various technical issues. The evaluator needs to consider whether processes are *observable* or the outcomes *measurable*. Are they defined in manners such that an instrument can be identified, or constructed, to answer the evaluation question? I have discussed earlier the various sources of data that might be used in an evaluation. You, as the evaluator, should consider these and determine whether there is the possibility of obtaining an appropriate instrument.

Also, how might the data be collected? What kind of data collection *schedule* would be required, and under what conditions should data be collected? Is there sufficient time for data to be collected? Is access to

the data resources possible? At this point, you may be quite flexible and expansive, and allow for multiple possibilities of data acquisition. Of course, if even in your most imaginative mode there are no ways of gathering the needed data, then the questions must be modified.

Think further—if the potential evaluation question requires a *comparison* between this program and an alternative, then you must consider whether alternative programs are present or possible. Does the evaluation require a design (more on this later) that necessitates random assignment of participants to one of two treatments (i.e., alternative programs)? If not, is there another program that could be used for comparison purposes? How might we gain the participation of those in the program that is not of primary interest?

A related technical issue to be considered is the potential receptivity to the evaluation within the organization. Does implementation of a potential data collection plan depend on a greater level of *cooperation and assistance* from staff than is likely to be extended? Will the required cooperation and access be available?

Time available is a technical issue; evaluations are conducted within specified time periods. The evaluator needs to develop a time line for what things will occur and in what order. I have already talked about time in the context of a program's development and ways in which this could influence when various evaluation questions could be asked and answered. A question might be appropriate in terms of the stage of development, but still could require more time than is available to answer it—that is, it may not be answerable within the period allotted, such as the contract year.

ETHICAL CONCERNS

There are ethical concerns associated with the selection of questions as well. In Section W, I provide some discussion of the various standards for program evaluation. Included in these standards are various ethical concerns related to the propriety involved in conducting an evaluation. These include the manner in which we deal with clients and other staff members. Evaluation must be conducted in a manner that does not intrude into the personal *rights of individuals and organizations*. Participants need to be informed about, and agree to participate in, the evaluation. Confidentiality must be ensured. If the evaluation needs to be conducted in a manner that requires the disclosure of the identity of the participant, or nature of his or her response, then this intrusion is inappropriate. So, consider carefully what will be required of those who participate in the evaluation. What can happen that is

bad or harmful? To the extent possible, seriously consider whether data collection might pose risks and harms to participants. Could the evaluation cause serious disruption to the program, or to the well-being of individuals?

Another ethical concern relates to you, the evaluator. As the evaluator, you should not consider the inclusion of a question that might potentially include some *conflict of interest* on your part. A conflict of interest may occur in the evaluation if you have a vested interest in the program to be evaluated. For example, do you have a loved one who funded the program, is employed by it, or participated in it? If so, then you might find yourself in rather difficult political circumstances and I would urge you to consider the extent to which you would be able to conduct the evaluation to the profession's standards (more on this in Section W). Granted, as I have noted, all evaluations are political, but be alert to questions that would be imprudent for you to address.

Finally, examine whether in considering particular questions, the nature of the data is likely to lead to *misleading evaluative information* or conclusions. Attempt to clarify these issues at the outset in order to avoid future problems. Be wary of the kinds of questions whose findings are particularly amenable to potential misuse (this topic is discussed in Section V).

POLITICAL FEASIBILITY

In Section H, I discussed the need to examine the organizational, community, and political contexts. I noted that evaluation is political in nature. Evaluation questions inspire a variety of strong views within the political context—views running in many different directions. Consider carefully the *political feasibility* of examining the questions. A funder who recently invested a significant amount of resources in a nascent program whose effectiveness is assumed, for instance, will likely be hard-pressed to view an evaluation as a systematic study rather than a public relations campaign. Likewise, a high-level university administrator might prefer to steer clear from questions about the ways in which the school's programs serve (or not) communities of need because such evidence might disrupt plans that they have been enacting. Answering these questions (or seeking answers to these questions) could be so politically disruptive that finding the answer is not possible. This does not mean that politically charged questions are beyond examination—rather, you should contemplate the extent to which the political environment allows the asking of the question.

➢ ***My Advice:*** Many potential evaluation questions are not readily answerable, and many more questions may have been posed than can possibly be addressed. Make a realistic appraisal of the feasibility of proceeding with the evaluation questions as projected by the sponsor and consider where appropriate a possible reduction in scope. Which evaluation questions, and how many are accomplishable? Consider reducing the projected data collection to what is both doable and most important. It is important that you, as the evaluator, make clear to stakeholders what is possible to do successfully.

RECAP—SECTION N

Are the Questions Evaluable?

- What Is the Nature and Stage of the Program?
- Are Appropriate Resources Available?
- What Is the Nature of the Question?
  - Question of little worth?
  - Answer really wanted?
- Can Standards Be Established?
- What Are the Technical Issues?
  - Potentially measurable?
  - Data collections possible?
  - Comparison available, if required?
  - Cooperation appropriate?
  - Sufficient time?
- Are There Ethical Concerns?
  - Rights of individuals?
  - Evaluator conflict of interest?
  - Potential for misleading information?
- What Is the Political Feasibility?

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

In addition to the questions outlined in the recap of this section, consider as well some of the underlying reasons why you may encounter evaluability issues. With respect to the nature of evaluation questions, for example, think about which spe-

cific stakeholders are interested in which questions and why. What is their motivation? For example, if Children's Trust and Family Matters are interested only in questions about children's health and school readiness statuses (i.e., program outcomes), then why might that be the case? If Amy Wilson is interested in questions about how parent leaders are recruited and how the curriculum is delivered across sites (i.e., program activities), again, why might that be? And if you engage parent leaders in the evaluation and they are interested in how RUPAS allocates human and financial resources to various activities (i.e., inputs), do you understand why?

Generally speaking, you must reflect on whether you have a firm grasp on the extent to which different stakeholder groups are motivated by the desire to improve existing program operations. Do they feel that the program is worth the investment and participation is beneficial, or are they motivated by some other factor and what would that be? Similarly, ask yourself this—to what extent is each stakeholder engaging in evaluation to satisfy external accountability requirements?

As for technical issues, consider whether there are gatekeepers with whom you should collaborate to work through possible challenges. For example, what role (if any) could Zoe and Carmen play? How would you connect with and engage them?

Further Reading

Davidson, E. J. (2015). Question-driven methods or method-driven questions?: How we limit what we learn by limiting what we ask. *Journal of MultiDisciplinary Evaluation, 11*(24), i–x. Retrieved from *http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/414*.

This article offers suggestions for articulating and focusing evaluation questions.

Liket, K. C., Rey-Garcia, M., & Maas, K. E. H. (2014). Why aren't evaluations working and what to do about it: A framework for negotiating meaningful evaluation in nonprofits. *American Journal of Evaluation, 35*(2), 171–188.

In this article, you will find a framework for organizing the relationships among evaluation questions, designs, and outcomes to be measured. It may be a helpful resource for engaging stakeholders in discussions about how to focus the evaluation.

Quick Reads

1. Ravneet Tiwana on Guiding Lights: Creating High-Quality Evaluation Questions and Engaging Stakeholders
   http://tinyurl.com/z33w2dx

2. Lori Wingate and Daniela Schroeter on Introducing the Evaluation Questions Checklist for Program Evaluation
   http://tinyurl.com/htk3pkl

SECTION

How Do We Plan a Process-Focused Evaluation?

Here we are. We have been working toward this—designing the evaluation. Arriving at this point required us to consider a number of different factors: the stakeholders and their values (Section E); the program and its logic (Sections G and I, respectively); the organizational, social, and political context (Section H); and the possible questions to be answered (Section J). In Sections K–M, we looked at what instrumentation might be used to answer possible questions—that is, we considered what data could be collected and what information we might obtain to answer the proposed questions. Finally, in Section N, we assessed all of these factors to determine whether answering these questions as the major evaluation focus was a realistic and doable undertaking.

However, you are still missing a few pieces related to the evaluation design and the evaluation plan. You need to know what *kind* of evaluation you will conduct. Is it focused on the manner in which the program is implemented—that is, is it process oriented? Or is it focused on the extent to which the program attained its stated goals—that is, is it outcome oriented? This last bit of preparation is an important step toward finalizing the evaluation plan and designing the evaluation. Let us take a brief moment to clarify the difference between each and then discuss how to go about a *process-focused* evaluation.

THE EVALUATION PLAN AND EVALUATION DESIGN

Let me make a distinction—perhaps an arbitrary one—between the evaluation "plan" and the evaluation "design." Typically, evaluators consider the "plan" as all of the procedural aspects that will make it possible. For example, the way in which you will collect data, or, more precisely, how will data be collected, in what form, and from whom? Also included in the plan are specifics as to how the data you collect will be analyzed along with considerations for how you will use the data to answer your questions. Finally, you will need to think about how to communicate your findings—that is, who will communicate to participants that the evaluation is taking place, when, and how; who will send reminders to participants about completing surveys if they are being used, when, and how often. The evaluation "design" focuses on issues of what comparisons will be made, how success will be determined, who gets picked to participate in the program and how, what instruments or measures will be used, and when measurement will take place. It consists of both process measures and outcome measures. In this section, we talk about the *process measures and their use in a process-focused evaluation*. Before doing so, however, it is appropriate to review what I mean by process measures.

Process Measures

As you recall, in Section A, after talking about the cook and his soup, I presented a description of purposes of evaluation, and showed the role of process measures in each of these types. However, you now know, after reading the discussion about theories of action in Section H, that the process is more complex. There are different kinds of processes. On the one hand, there are *administrative processes,* which involve the *program inputs* (or resources) needed to implement the program. When you start collecting evaluation data, you will want to examine whether these various inputs are, in fact, present. On the other hand, there are *implementation processes,* which involve *program activities* and *outputs*—the things that are supposed to happen within the program, such as workshops, tutoring, and field trips, and their frequency. Evaluation data that you collect about activities and outputs will focus on the extent to which they are being offered as intended. There is another kind of more complex process to be examined as well. Based on the program's theory of action (as depicted in the logic model), there is a belief that some kinds of mechanisms are operative. I refer here to a process that connects a program activity with a specified outcome. When, for example, one teaches a unit in a mathematics course, why does one have an

expectation that certain understandings will be achieved? What goes on, or is supposed to go on, in a student's head in order for that learning to occur? I refer to this aspect of process as *program mechanisms*.

As you might intuit, these various kinds of processes have different applicability to the purposes of evaluation. This is shown in Table O.1. As you can see, formative evaluation is primarily concerned with administrative and implementation processes. The focus here is specifically on the activities that are offered, the inputs that support them, and the factors that affect how services are delivered. Summative evaluation is primarily concerned with program mechanisms and outcomes. In a summative evaluation, it is assumed that the program has undergone extensive formative evaluation and the program inputs and activities are now in place. Thus, in a summative evaluation, which generally seeks an understanding of causality, the evaluator wants to understand whether the program is producing the outcomes for which it was designed. Summary formative evaluation may include a focus on program inputs and activities *as well as* program mechanisms. In that instance, the program inputs and activities, respectively, are examined so that resources can be redistributed if necessary and modifications can occur if services are not being implemented as intended. Program mechanisms are examined not so much for issues of causality, but for understanding how inputs and activities should be adjusted and strengthened so that they eventually produce evidence that supports claims about causality. As you can see, program mechanisms sit at the intersection of process and outcome evaluations. We address it from a process-oriented perspective here and expand on it from an outcomes point of view in Section P.

Now that we have reexamined the types of evaluation and clarified the three kinds of process variables, let me discuss how they might be considered in the evaluation plan. As noted, there are administrative processes, implementation processes, and program mechanisms. Let us consider each of these in turn.

ADMINISTRATIVE PROCESSES

I discussed various program inputs in Section G, when we talked about "What is a program?" In that section, I mentioned people (personnel), materials, facilities, financial resources, and clients as examples of inputs. Your job in developing the evaluation plan is to consider how you will go about determining that these resources are in place and to articulate these potential steps, in general terms, in the plan.

Many of the more specific details of gathering these data will not be included in the plan, but will be determined when you subsequently

TABLE O.1. Purposes and Types of Evaluation

Purpose of evaluation	Process			Process/ outcome	Outcome		Program audiences
	Administrative	Implementation					
	Inputs	Activities	Outputs	Program mechanisms	Short and medium term	End of evaluation	
Formative	×	×	×				Program staff
Summary formative	× (limited)	×	×	×	×	× (limited)	Program staff, program designer, stakeholders
Summative				×	×	×	External audience, stakeholders

examine the implementation of the program. However, in this section of the book, I provide an indication of some of the things that you need to look for. Consider the following as *general guidelines* for examining administrative process variables, specifically program inputs.

Personnel

Now let us jointly think about personnel—primarily the program staff. Here are some questions: Have the right *number of staff* in the right categories been hired? Have there been personnel turnovers or understaffing due to a failure to hire? Does the staff present at the site have the *appropriate qualifications*? It may well be that, for the program to be effective, staff need to have particular training or unique skills. They may need to have certain cultural understandings or sensitivities. You will want to consider whether the staff is capable of taking the actions required of them in conducting the program.

Furthermore, you will want to know whether the personnel have an *understanding of the program,* an understanding of the program's goals, and facility in implementing the required program activities. *Attitudes of staff* are also important because this might constitute one of the reasons for the program elements not being put in place. You will want to consider whether the staff is working together in an effective manner. Staff might not be willing to make changes for a variety of reasons: they might hold on to old ideas; there might be conflicts with other staff or administration; or, perhaps, they might be frustrated by trivial obstacles that occur in almost any new program.

Many of the kinds of questions about personnel appropriateness can be answered by an examination of the program's personnel files. Who are the staff? You will need to look for information on their academic training and experience. However, some personnel attributes are simply not found on paper and will require different procedures. Interviews and observations are important tools for gathering data related to personnel and other relevant components of the program. An important area to be covered in interviews is the perceptions of program staff (and the perceptions of others as well). Do they have appropriate feelings about participating in the program? Are they able to relate well to program clients?

Note that we are not focusing on the quality of staff's work nor are we trying to arrive at value statements about staff performance. We are not evaluating staff—that is personnel evaluation. Here, we are focusing on program evaluation and whether the program has the sufficient human resources to go about its work. Said differently, we might consider this component of the process evaluation as a "fact-finding mission." What

staff is in place? Do they have the appropriate skills? Is the number of staff available sufficient given what the program aims to accomplish?

Materials

Programs typically do not consist of personnel alone. Personnel employ methods that include a variety of materials. Materials include such things as computers, books, handbooks, manuals, and so on. You will need to consider whether they are all there. Is there enough of them? Did they *arrive on time*? Note, however, that "on time" is a relative notion. Hopefully, materials will arrive prior to the start of the program, so that staff can become acquainted with them, and be ready for full operation when the program starts. Sometimes all materials are not needed at the start of the program, so it will be necessary to compare the time line of the program to when the materials are scheduled to be employed. You, as the evaluator, will need to determine the extent to which late implementation of the materials impacts the program, and the extent to which the conditions are subsequently remedied.

Facilities

What about *facilities*? Are they as anticipated? The program, to operate properly, requires a certain availability of space, and it is important to determine whether that space or comparable space was provided. For example, you might notice that the program is buried in a trailer too small in which to function. The program might not be able to operate properly if there is insufficient room to conduct the program activities as anticipated. Furthermore, while the space may be appropriate in square footage, there might have been special requirements related to configuration. You will want to note whether the space meets the needs of the anticipated program activities. Furthermore, administrative tools are also important. Consider, for example, whether there is an active, ongoing record-keeping system and if there are avenues for the sharing of this information within the program. Without an active information flow, program administrators will be unable to note deficiencies and seek to correct them.

Financial Resources

Many of the deficiencies in implementation of program activities may be related to a lack of appropriate financial resources. It may well be that the budget contemplated in the plan was insufficient for meeting the needs as stated—this is a plan deficiency that will need to be attended to. The more important question is whether the personnel, dollar, and

materials allocation actually provided and used corresponds with what was envisaged in the program plan. In either case, this information will need to be examined so that program staff can deal with the issue. These resource issues are likely to be already known by program staff, but not necessarily by all primary stakeholders.

Program Clients

Programs have clients—individuals who are believed to be the beneficiaries of those programs. The program was conceived with particular clients in mind—that is, when thinking about what kinds of activities, what kind of staff, what kind of materials, and how activities should be structured, these elements were designed to benefit specific program recipients. Thus, it is imperative that you examine whether the "who" that the program is addressing is the "who" that had been anticipated. Are they *the right clients*—the ones intended? Looking at the makeup of the client population is important in knowing whether the program has been implemented as designed.

What is it about the clients of the program that is important? First, you will need to determine whether their number corresponds to what is anticipated, as specified in the program plan. If there are too many clients, for example, that might be a cause for reduced effectiveness. In addition to number of clients, certain presupposed clients with particular needs, particular characteristics, particular prior training or educational level, and so on had been anticipated. You will need to determine whether these kinds of clients were actually served. Consider, for example, an after-school enrichment program for struggling students that was intended to improve student outcomes. However, there was an open sign-up that offered an opportunity to anyone who wished to enroll and attracted self-selecting, high-achieving students. Because of this modification of the target client group, it would be anticipated that the program would not perform as expected.

But beyond the question of appropriate clients, I want you to consider another issue. Are these *clients being retained*? Are clients completing the segments of the program on an appropriate schedule? Are they dropping out of the program en route? All of these are important client-related process issues that you will need to examine.

IMPLEMENTATION PROCESSES

We have carefully considered the resources that are necessary to operate the program as part of our effort to evaluate administrative processes. Specifically, we considered the availability and adequacy of

these inputs. Now, we seek to understand the role that they play in the implementation of program activities and their outputs.

Examining the implementation of such activities requires us to "open the box" and to take a careful look at what's happening inside that "box" (i.e., inside the program). Thus, process evaluations that take implementation as their focus deal squarely with the issue of determining what activities (or services) were delivered, which resources were used to enact those activities, how were they delivered, by whom, what steps were involved, and what numerable returns can be observed.

Let's take the tutoring program that we discussed in Section I as an example. The program called for several activities, each to be completed by different groups of people. If we were to examine the implementation of the program, then what sorts of things might we focus on? To start, perhaps we aim to learn whether all of the activities as depicted in the logic model were being delivered. If not, then which ones did program staff decide not to offer and why? Did the program encounter undue financial or political challenges, for instance? And how representative were such difficulties of prior year experiences?

Now that you have determined which activities are unavailable, you will want to more carefully examine those that are indeed offered. Of those activities, learn about the intricacies involved in implementing them well and the factors that could prevent staff from doing so. Looking at the program plan, you will need to examine the *intended activities* and the *sequence* in which they occur. Questions you will need to answer include: Did these get implemented in the appropriate order? Did these get implemented in accordance with the program's time line? Take, for example, the community walk that the tutoring program requires of its student employees. Make it a point to learn who is expected to complete this activity. In this case, it is undergraduate tutors. Then find out if this happened. If yes, what made it possible? If not, why not? What were the challenges encountered?

In this process, you might uncover a number of different systems-, community-, or program-related factors that staff had to work with, through, and around. It is entirely possible, for example, that staff had to forgo the program's partnership with the transportation company that had been busing students to the same three or four communities these past 8 years for a more affordable vendor. You might also learn that 67% of the administrative staff were all new and had little experience implementing this aspect of the program. Thus, some students were sent to incorrect communities or some schools received incorrect information and were not adequately prepared to support the program. Alternatively, you might come to discover that partnering schools and other community organizations donated time by doing the walk

with students, which alleviated the program's staffing challenges and enabled tutors to have an adequate orientation session. These insights might encourage you to probe for a better understanding of the sorts of efforts and resources that are needed to hold a successful community walk. For instance, how many people are needed to plan, coordinate, and execute this single program activity? What steps must staff undergo and how far in advance must planning take place?

Likewise, you will need to understand the tangible results of offering such an activity—that is, you need to have a pulse on the activity's outputs—those things that can be counted. For instance, how many of the undergraduate tutors completed the community walk? If less than 100% participated, then why might that be? Furthermore, how long was the community walk supposed to take? If 2 hours, was the full 120 minutes allocated and used? If not, why and what took place instead? If there were specific places that tutors were supposed to see, particular people with whom they were supposed to interact, exercises that they needed to complete during the walk, then to what extent did these happen? How do the outputs in these areas compare with previous iterations of the walk?

You see, examining implementation processes calls for explicit understandings about all of the activities that the program seeks to provide. (Note that we have only begun to scratch the surface with the community walk as *one of several* activities that the tutoring program offers.) Equally important in this endeavor are the audience of each activity and what it takes to get it done. A thoughtfully conducted implementation evaluation also requires insight into how the program's work had been done *previously* and whether it is *now* being done as depicted on the logic model. You will notice that there is iterative comparison between what's intended, what's implemented, and how it has evolved through time. All of this information, together, allows you—the evaluator—to arrive at insights about the program's implementation successes and challenges. Findings about deviations from the logic model will reveal at least two things: (1) ways in which program activities, and the program itself, were not implemented with fidelity; and (2) the staff's creative response to barriers encountered that are outside of their control.

PROGRAM MECHANISMS

The previous discussion referred to an examination of program inputs, activities, and outputs—that is, administrative and implementation processes. But, as previously noted, processes are more complex than

that. They also refer to whether such elements (even if seemingly implemented properly) are achieving the desired results—the relevant short-term outcomes. This is perhaps the most difficult aspect of a process evaluation. One of the challenges lies in understanding the extent to which the stated activities and the intensity with which they are offered are sufficient in triggering desired changes in the short and long terms.

Well, how will you do it? You will need to reexamine the program theory to see what kinds of activity–outcome relationships are reflected in the logic model that you might have constructed with clients (see Section I). That logic model helped to define the rationale for why different activities were taking place. Discrete activities were employed, or put in place, to achieve particular short-term ends (e.g., certain knowledge, understandings, skills, attitudes, or behavior changes). And these, in turn, were designed to lead to the ability to engage in further activities and trigger other outcomes. In our sample tutoring program, undergraduate tutors are expected to complete an orientation in the community where they serve. If they do not complete this activity, then they are unable to engage in the next one—which is to offer tutoring. Likewise, if they do not offer tutoring, then the subsequent related outcomes will not be observed. The likelihood of the mechanism being in place for further evaluation, here, is very low. The aim in this phase of examining program mechanisms is to determine the *potential existence* of the connection between program activity and short-term outcomes by carefully considering the logic that is supposed to hold up the program in the first place.

Next, you will need to determine whether the "dosage" (i.e., the frequency and intensity) of the delivered activity as experienced by participants is sufficient to lead to short-term outcomes. Accomplishing this will require access to materials that document participants' engagement in such activities. In our example, this information might come to you in the form of logs that detail who came to the orientation and the times that they signed into and out of the event. Clearly defined and measurable short-term outcomes are also needed.

The logic model in Section I does not articulate short-term outcomes explicitly related to the orientation as a program activity. Rather, three long-term outcomes are indicated. Among them is "Tutors will feel motivated to work with low-income communities." This outcome could be measured in a number of different ways. You learned from Section K that we can obtain self-report data from tutors using surveys. We can also survey those who know the tutors about whether they have been seen working in disadvantaged communities in other capacities. This would be another source of evidence for such motiva-

tion. Alternatively, from Section L you learned that we can also use interview or focus group methods to collect such data. That would suffice here as well.

Once such data have been collected, you can then go about systematically comparing tutors who completed orientation against those who did not to determine whether there are differences in levels of motivation. Likewise, comparisons between tutors who completed the full orientation could also be made against those who did only a partial session. These findings will help you to begin understanding whether a one-time orientation helps tutors to feel the motivation that program staff anticipate. Such insights can also suggest avenues for improving the program and its logic model.

We noted earlier, for example, that the tutoring program's logic model does not indicate short-term outcomes related to orientation. This observation begs the question of whether one is needed. Did we by chance overlook other important mechanisms related to motivation that should be included? Should we account for various intrinsic (e.g., tutors seeing reflections of their home communities in the ones where they work, the opportunity to do good) and extrinsic (e.g., financial incentives) factors that drive some tutors and not others? Likewise, should we consider other desirable unintended consequences, such as tutors learning something new about these communities that they did not know previously? Or perhaps some of the tutors' previously held assumptions about these places are being challenged?

Process evaluation focused on program mechanisms seeks to examine the existence and strength of the kinds of linkages described here. You will want to know: Did the anticipated *linkages* in the program's logic model hold up? Did they *make sense* in practice? Did conducting a particular activity lead to certain understandings, which, in turn, allowed another activity to take place efficiently?

QUESTION: How do you, as the evaluator, begin to understand this?

ANSWER: With great difficulty. It requires selecting particular linkages in the logic model and gathering data about the results and consequences of an activity to determine whether the specific contribution had been achieved. Clearly, there are far too many conceptual linkages within the logic model to examine all of them. Moreover, some may be so complex that measuring attainment would be extremely difficult. Thus, you, as the evaluator, might only select several of them for examination based on their importance within the logic model. In doing this, consider which activities, and which linkages, are presumed to be most important.

RECAP—SECTION O

Process Measures

- *Administrative Processes*
  - Personnel
  - Materials
  - Facilities
  - Financial resources
  - Clients
- *Implementation Processes*
  - Intended activities
  - Sequence of activities
  - Factors that contribute to and hinder implementation fidelity
- *Program Mechanisms*
  - Program theory making sense
  - Potential existence of linkages
  - Sufficient intensity of activities
  - Preliminary between- and within-group comparisons

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

We have explored and considered a number of different contextual factors that not only have helped us to better understand the RUPAS program but also to evaluate it in a meaningful manner. Let us assume, for a moment, that your primary intended users are decision makers at Family Matters (FM) and they are commissioning a 2-year evaluation intended to shed light on the RUPAS program's inner workings—that is, its processes. They hope to use evaluation findings to establish some guidelines for "good practices" as they continue to grow the program and seek external support from philanthropic organizations. How might we approach such a process-focused evaluation?

Perhaps we start with the evaluation questions and the program logic model as orienting tools. As we study them, we might ask: What connections exist between program resources, activities, and outputs as described by stakeholders? How do they compare with what is reflected in the logic model itself? And, how well do these program elements align? Let's think about this. Would degree of alignment influence how you go about studying program processes and implementation of activities? If so, in what ways? For example, how would you proceed if the manner in which FM describes the RUPAS program is completely different from how Amy Wilson and her team are implementing it? What if FM anticipated that each site

would host weekly Parent Leader meetings, but in reality that is happening on a monthly basis—that is, what would you do if you had a fidelity of implementation issue on your hands?

If another stakeholder commissioned the evaluation—Amy Wilson, for instance—with similar conditions in place, would you change the ways in which you would navigate the study? Why and how?

Consider, as well, the views that are represented at this stage (the design stage) of the evaluation. Whose are they, what are they, and how might they influence the evaluation's trajectory?

Further Reading

After the program is put in place, you will proceed to examine various process variables. In essence, you will then be determining whether the program was implemented properly. The notion of *implementation evaluation* is implicit within this section—it is an important topic. I believe that examining some additional readings would be beneficial. Here are some of my favorites:

Judd, C. M. (1987). Combining process and outcome evaluation. *New Directions for Program Evaluation, 1*(35), 23–41.

This article describes approaches for conducting process- and outcome-focused evaluations simultaneously, which is often required in real-world evaluation practice.

King, J., Morris, L., & Fitzgibbon, C. (1987). *How to assess program implementation*. Thousand Oaks, CA: SAGE.

This short monograph describes the way that process elements are used in examining whether programs have been properly implemented.

Mathison, S. (Ed.). (2005). Process evaluation. In *Encyclopedia of evaluation* (pp. 327–328). Thousand Oaks, CA: SAGE.

Consult this encyclopedia entry for an introduction to what process evaluation is.

Perez, J. L., & Yerena, A. (2016). Evaluating the policy–practice gap in a transitional housing program: An innovation in process evaluation. *American Journal of Evaluation*. [Epub ahead of print]

This article offers an interesting case and perspective for considering and conducting process evaluation of a particular social program that seeks to address housing needs of vulnerable communities.

Saunders, R. P., Evans, M. H., & Joshi, P. (2005). Developing a process-evaluation plan for assessing health promotion program implementation: A how-to guide. *Health Promotion Practices, 6*, 134–147.

This article provides a framework for developing a comprehensive process evaluation plan using the case study of a health promotion intervention.

Stufflebeam, D. (2005). CIPP model (context, input, process, product). In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 61–65). Thousand Oaks, CA: SAGE.

For a discussion of the CIPP model—one of the earlier evaluation models to explicitly discuss inputs and processes in evaluations—read this encyclopedia entry.

Quick Reads

1. Clara Hagens on Guidance on Monitoring and Evaluation
   http://tinyurl.com/hobz6bg

2. Mindy Hightower King and Courtney Brown on a Framework for Developing High Quality Performance Measurement Systems of Evaluation
   http://tinyurl.com/hnkmcjm

3. Adam Kessler on Why Evaluations Fail: The Importance of Good Monitoring
   http://tinyurl.com/zktqnlu

4. Monica Oliver and Krista Collins on Creating an Index for Measuring Fidelity of Implementation
   http://tinyurl.com/j8r4zwr

SECTION

P

How Do We Plan an Outcome-Focused Evaluation?

In Section O, I discuss ways in which we might study and evaluate program processes. In this section, I discuss how we address questions about program outcomes. What do I mean by outcomes? At its core, outcomes are what we anticipate seeing in program beneficiaries because they completed some subset of program activities or participated in the program in whole. From Figures I.1 and I.2, and likewise Table O.1, outcomes are visually located on the right side of the logic model and table, respectively. Generally speaking, outcomes can be challenging to measure, especially when compared with program processes. The manner in which they are defined plays a role in determining measurability as well. So we begin our discussion by considering the issue of outcome definition, and then about how to answer evaluation questions related to the measurement of outcomes. We also address how to better understand the mechanisms (remember this concept from Section O?) that contribute to the attainment of such results.

OUTCOME DEFINITION

I mentioned in earlier sections of this book that there exists several different types of outcomes. Social programs in general seek results in three areas: understanding (i.e., learning), attitude, and behavior (or skills). The program's scope and the field in which it is located defines, in part, what is measured. Academic support programs, for example,

often emphasize learning (or the acquisition of knowledge). Health-focused programs tend to be interested in behaviors that promote well-being. Rehabilitation programs address learning and behavior as well, but also invest in changing participants' attitudes. That said, it is often the case that programs value all outcome areas but resources are finite. Thus, programs will allocate varying amounts of resources to each area to ensure that their daily operational goals are in line with long-term organizational goals. As the evaluator, you need to understand this and confirm the extent to which it applies to the program that you are evaluating.

Outcomes can also be considered in terms of time. Recall that outcomes on the logic model in Section I are divided into short-, intermediate-, or long-term categories. This is an important distinction, particularly in the measurement context, because some programs may have been designed with the participant's growth and development in mind. Thus, outcomes might be laid out over the course of several years on the logic model—that is, short-term outcomes might correspond with results expected at Year 1, intermediate-term outcomes with Years 2 and 3, and long-term outcomes with Years 4 and beyond. While articulating outcomes in this manner seems intuitive, it would behoove the evaluator to be prepared to work outside of this structure since it is entirely possible for the program's outcomes to be organized differently.

While we are on the issue of understanding outcomes, here is one more note related to their definition—outcomes, as defined by funders and program staff, can appear relatively generic on the surface. As I have suggested, academic support programs are likely interested in what and how well students learn. However, the manner in which academic content and students' performance are operationalized within various programs is distinct. Fourth-grade students' ability to solve algebra problems, for example, is different from welders' ability to artfully fuse two pieces of metal together. Learning and demonstration of mastery are of interest in both cases, but, as you can see, the manner in which each is defined—and, therefore, the kinds of evidence that you would seek to determine attainment—will in turn be different. Context is key.

THE EVALUATION PLAN AND EVALUATION DESIGN

In Section O, I used the term "design" to refer to the description of the way in which the evaluation could be conducted when the focus is on processes. This included the selection of individuals and instruments and the way in which data could be collected and analyzed. Designs for outcome measures can be complicated, but they all get at the same

issue—to what extent did the program reach its stated goals—that is, did it work? Researchers talk about a variety of sophisticated ways for designing studies involving the measurement of outcomes and I think it is helpful to simplify what to look for in developing evaluation designs.

A friend and colleague of mine, Michael Patton, called my attention to a nice quote from Rudyard Kipling about the six honest serving men of evaluation:

> I keep six honest serving men.
> They taught me all I knew:
> Their names are What and Why and When
> And How and Where and Who.

Let us consider these generic questions as the basis for describing what is included in an outcome-oriented evaluation design. I have left off the "where" for convenience, since typically it would occur on-site.

I choose to start with *why*. The evaluation design is based on the question of "why." Why are we doing this evaluation? This is very important as it is the reason I have spent so much time talking with you about what the question is.

Next, consider *what*. In this case, I let the "what" refer to the data. What data are potentially available for answering the question? What is the concept to be measured? What is the manner in which these data will be acquired? Will we give a test? Perhaps administer a questionnaire. Possibly, observations. It could be that we will conduct interviews. The means of data acquisition, along with the concept to be measured, need to be specified to ensure that we have the data needed to answer the question.

If there is to be some data collection—that is, acquiring data—*who* (or *whom* if we are to be grammatically correct) will it be obtained from? Will we get data only from clients of the program? Will we get data from other groups that are similar to those in the program being evaluated? Perhaps even from groups that have been selected in some particular way. Will we ask for information from those who staff and administer the program? If we are concerned about the impact of the program on the community, will we gather data from community stakeholders?

Depending on the question that is asked there are a variety of times *when* data might be acquired. Do we want to know clients' initial understandings of information to be presented within the program? Thus, data collection would begin when clients begin the program. Do we want to know what progress clients are making? Thus, data would be collected at multiple intervals throughout the program. Consider the necessary time schedule for measuring that progress. Are we con-

cerned about the outcomes of the program—namely, what do clients know at the end of the program? These are "when" questions.

At this point I ask you to consider the *how*. How will the collected data be aggregated—that is, will we summarize it? Will we submit it to some statistical analysis? Will we provide a very detailed qualitative description? (Evaluators sometimes refer to this as "thick description.") Also, how will we decide what these aggregated or described data mean? For example, can we discern that the question has been answered positively? This is the valuing component and standards need to be established as part of the evaluation design that enable us to make these judgments. (More on this in Section T.)

RECAP 1—SECTION P

Elements of an Evaluation Design

- WHY?—To answer questions.
- WHAT?—What data are to be acquired and the manner of acquisition?
- WHO?—Who will the data be acquired from?
- WHEN?—When will data collection take place?
- HOW?—How will data be analyzed and valued?

AN EXERCISE TO ASSIST US

I think it might be helpful to consider some sample evaluation questions, and as we do, I ask you to examine the implications of each question for the "what," "who," "when," and "how" issues discussed above. In order to do this, let us consider the Undergraduate Tutor Training (UTT) Program that I referenced in Section O since it provides a context for our subsequent discussion.

DESCRIPTION OF THE UNDERGRADUATE TUTOR TRAINING PROGRAM

The Undergraduate Tutor Training (UTT) Program is located on a university campus in a bustling urban college town in Southern California. The program has partnerships with underresourced elementary and middle schools throughout the city. These schools enroll predominantly minority students of African American and Latino backgrounds. Space does not allow for the full description

of the school context, but suffice it to say that it is characterized by students with rich cultural experiences and diverse linguistic talents. At the same time, these schools also suffer from a high transiency rate and high staff turnover. Moreover, there is a substantial resource issue because the program is funded by the university and the budget landscape is unpredictable each year. The program develops its tutor training curriculum and works with school staff to schedule days and times when its tutors can visit each site. The curriculum addresses how to teach reading and mathematics to fourth- and eighth-grade students, respectively. It also covers classroom management approaches and local and federal educational policies along with ways in which tutors can be mindful of how students can contribute to their own learning. Tutors rely solely on books and homegrown learning activities and worksheets to help their students. Student test scores across partner schools have consistently been among the lowest in the school district.

The kinds of questions that we might ask can be thought of as either "causal" or "descriptive." Causality implies that the linkage between outcomes attained and the intervention (the program) can be demonstrated. Descriptive means that results or attainments are presented, but we do not know whether other factors may have been partially responsible for obtaining the results. It is exceedingly difficult to prove causality—that is, I repeat, that a program intervention was *solely* responsible for changes in the outcome. Consider for a moment the following simplistic question:

QUESTION 1: Does the UTT Program work?

Well, what do I mean by "work"? One must presume that the intent of the question is that the outcomes attained are meritorious or worthy and the UTT Program was responsible for this desired outcome—namely, that the program *caused* the positive result. I return to the causality issue, but let us first consider the issue of the positive outcomes. We want to know whether the resulting outcomes were valuable. How do we know what is valuable? We might look at student scores and notice that they have increased, and so, this question might be rephrased.

QUESTION 2: Does the UTT Program produce improved student outcomes?

If the outcome scores are higher, does this prove that these improved outcomes were due to the UTT Program? Not really. Consider what else

might have led to changes in outcomes. Researchers talk about various factors that weaken an evaluation design. These are factors that in some way make less certain the intent of the design. One of these is referred to as *maturation*. Maturation is the notion that there is a natural growth in the participants of programs. They are getting older and they may be socially and psychologically maturing—this improved score might be related to this maturation. There may also be other events occurring in a student's life or in the student's social context generally that were providing knowledge and insights that might have partially accounted for the change in outcomes. Researchers and evaluators refer to this factor as *history*—that is, students at partner schools might be participating in other academic activities that in some way enhances their learning. Or perhaps, discussions between program staff and school administrators provided insights into some aspects of the curriculum. A third factor that might weaken the case for arguments of success is referred to by researchers as the *testing effect*. Simply stated, students become more test-wise. This is particularly true when they have been given a pretest.

TOWARD STRONGER CAUSAL MODELS

So, how might we compensate for these factors that potentially weaken the simplistic design that we intended? One way is that we could have two *comparison groups*, with the idea being that the same maturation, history, and testing factors would impact each of them equally. Students at partner schools would receive support from tutors in the UTT Program and the other group would receive tutoring from another program. (Note that in some instances, comparison groups receive no tutoring at all. That is not possible in the partner schools.) However, in this example, we could have tutoring available in some classrooms at some partner schools but not in others.

What is the danger here for not being able to make causal statements? Obviously, the problem is that not all classrooms—and not all schools—are the same. Classes at some partner schools might have more high-achieving students (or lower-achieving students). One solution is to give each of the classrooms within partner schools a pretest in which we attempt to determine their initial status—this is called establishing a *baseline*—and then try to control for differences statistically. In this design, we could have a *comparison group* that may not be exactly equivalent, so we take a baseline measure to see the initial differences in the group (pretest) and subsequently give a test at the end (posttest—what is called a *pretest–posttest comparison group design*). The important point to note here is that we are studying schools and

classrooms *as they exist*. We are not changing anything about the school and classroom settings or about how the program is delivered in these environments. Now let us look at our focus questions for design analysis. Consider the following for an answer to Question 2 using a pretest–posttest comparison group design.

Question 2

What data are needed?

- Academic achievement measures
- Attained through school testing

Who will the data be obtained from?

- Students in partner school classrooms who receive tutoring from the UTT Program
- Students in partner school classrooms who receive tutoring from an alternate source

When will the data be acquired?

- Beginning of school year for all classrooms at all partner schools
- End of school year for all classrooms at all partner schools

How will these data be analyzed?

- Statistical analysis of the differences

This is a better solution, but still not a perfect one. The small number of classrooms would make statistical controls difficult. Well then, we could more carefully compare outcomes of partner school classrooms that receive tutoring from the UTT Program against outcomes of nonpartner schools that do not have tutoring at all—that is, these nonpartner schools could serve as a comparison group. I think you see what the difficulties might be here. How will we get schools that are exactly comparable? And if we could find them, imagine the difficulty of getting them to agree to have their students tested without the potential benefits of receiving tutoring or some other compensation. Moreover, doesn't the fact that they are different schools, no matter what comparable qualities they share, ensure that there are some characteristics that differ?

Let's think some more, and think about what would be a better alternative for suggesting that the UTT Program (and *only* that program) led to the improved outcomes. Researchers suggest that a design called "the posttest-only control group design" provides the strongest assurance of causality. In this design, there is the necessity for random selection of participants (denoted by "R," below)—that is, each of the students at a grade level in partner schools would have an equal opportunity to be selected to be tutored by program tutors or other tutors. Tutoring sessions would be filled in that manner and outcomes would subsequently be measured (denoted by "M," below). This design would be called a *randomized controlled trial*. This provides a better answer to Question 2. In essence, the previous "working people" description is applicable with the addition of randomization, as shown below:

R—UTT Program Classrooms → M
R—Regular Classrooms → M

In this depiction, randomization takes place (as designated by the capital "R"); one group receives tutoring from tutors who were trained by the UTT Program and the regular classrooms from other tutoring programs. Following instructions, each group is measured on the same outcome indicator. Using this procedure we would have eliminated the design threat of maturity (students in each group *presumably* mature in a comparable way), and we partially eliminate the design threat of testing since no pretest is given. However, in this situation we continue to potentially face the evaluation design threat of "history," as mentioned earlier. Will students and teachers in the "control" (or regular) classrooms be aware of what is happening in the UTT Program classrooms and possibly incorporate aspects of it into their own settings? It does not appear then that by focusing solely on partner schools, even with randomization, we are able to definitively prescribe causality to the UTT Program undertaking.

However, most of the above-mentioned design possibilities are nonetheless potentially valuable in providing insight into what is happening at partner schools and provide *a strong case for suggesting causality*.

Another kind of question seeks to examine progress over time (or trends in specific outcomes), especially after a clear and definable change in the environment has taken place. Changes in policy and introduction of new tools and systems (e.g., curriculum, technology) are some examples. This is referred to by researchers as an *interrupted time-series design*. When this approach is used with a comparison group it provides causal insights. Consider the following:

QUESTION 3: Are reading scores among partner school students improving?

This kind of question is best answered when existing data are available *prior* to the implementation of the program, which it would be in this case, and *immediately after* the program began. The question is addressed by examining the trend changes over time. Thus, in addition to data that can be tracked, you also need a specific program start date. See, for example, Figure P.1.

As seen in the graph, scores were on the decline prior to the start of the program. I have shown a solid trend line depicting students' attainment of reading skills prior to the start of the program. (In essence, this is a line that best fits the assembled data—meaning, it fits best.) The dotted line depicted indicates the expected continuance of the trend based on being tutored by college-age students not trained by the UTT Program (i.e., scores of students in regular classrooms). Then I considered changes in annual reading scores among students in UTT Program classrooms after the implementation of the program. This is depicted with a dashed line in the graph. Showing the data set would provide a visual indication of the possible impact occasioned by the

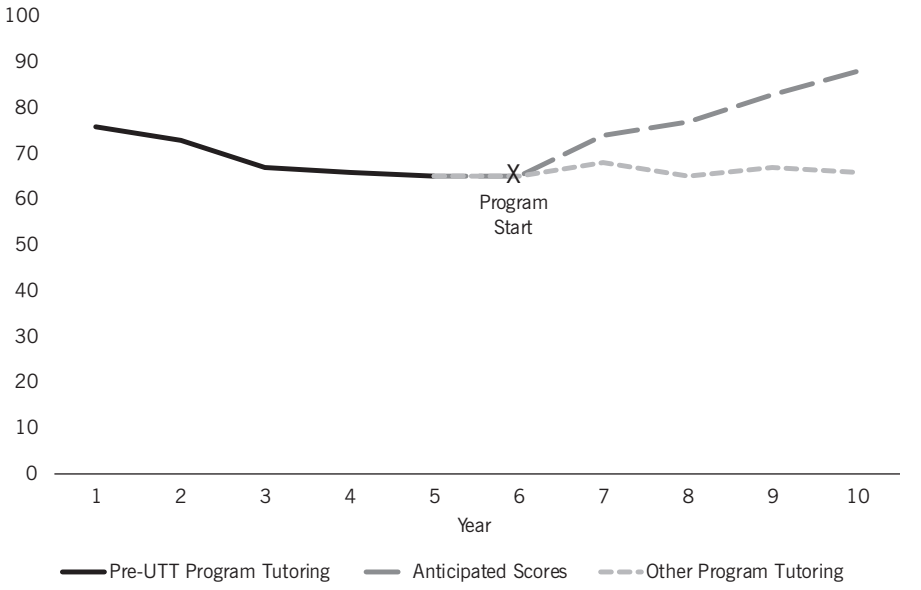

FIGURE P.1. Interrupted time-series data.

implementation of tutoring by the UTT Program versus the comparison program. Statistical procedures are available for providing a quantitative analysis of such data.

Now, why don't you fill in the chart for Question 3?

QUESTION 3

What data are needed?

- Your thoughts?

Who will the data be obtained from?

- Your thoughts?

When will the data be acquired?

- Your thoughts?

How will these data be analyzed?

- Your thoughts?
  (See end of section for answers.)

An offshoot of any of these questions or of many of the descriptive questions that follow might address the differential impact on different subgroups (e.g., males vs. females, different ethnic groups). This, of course, would require a focus on data about the subgroup and might have a variety of "what," "when," and "how" answers depending on the specification of the question.

Let me provide a *final* note on the issue of causal designs. Michael Scriven, a noted evaluator/philosopher, has suggested that causation can be directly seen in numerous other ways. He advocates for the systematic application of logical thinking, involving the consideration of alternatives, specification of likely causes, and carefully eliminating those that do not ring true (my words, not his). See the article by Michael in the Further Reading for this section.

DESCRIPTIVE DESIGNS

Another potential question that might be asked of the UTT Program is the following:

QUESTION 4: Is a greater percentage of students performing at grade level or above in reading at the conclusion of 1 year of the program compared with the prior year?

This is an infinitely easier question to answer. The intent of the question is not to show causality but to *describe*. As differentiated from causal questions, the design to answer this question and others are *descriptive designs*—they describe status.

Now, back to considering this question. In this instance, we have established a *standard* for judging whether the outcome is satisfactory. An examination of the percentage of students who were performing at grade level at the end of the prior year is compared with the percentage attaining grade-level or above status at the end of the program year. Similar questions that involve standards might stipulate different expectations related perhaps to the mean grade-level performance or to other specifically stated standards. For example, did students know more about technology at the end of the year than at the beginning? Or did 80% of the students learn how to access and use the Internet?

Read Question 4 and provide answers to our "working people" questions below.

QUESTION 4

What data are needed?

- Your thoughts?

Who will the data be obtained from?

- Your thoughts?

When will the data be acquired?

- Your thoughts?

How will these data be analyzed?

- Your thoughts?
  (See end of section for answers.)

The following is another situation entailing questions that we might want to ask about the UTT Program. This may also be considered as a descriptive design.

QUESTION 5: Do students, parents, teachers, and administrators view that the program is performing in a satisfactory manner?

If students are dissatisfied with the program, this is important to know. Perhaps the heavy technical engagement in the learning of mathematics is making students more math phobic. Do parents believe that their child is benefiting from the program? Do they feel that their child is more (or perhaps less) interested in school? Similar questions could be asked about teachers and administrators. In essence, this is a *one-time snapshot* of some outcome of interest. For this question, you would want to employ surveys, interviews, or possibly, focus groups. Obviously, not everyone would be interviewed or participate in a focus group, and so careful selections would need to be made.

QUESTION 5

What data are needed?

- Your thoughts?

Who will the data be obtained from?

- Your thoughts?

When will the data be acquired?

- Your thoughts?

How will these data be analyzed?

- Your thoughts?
  (See end of section for answers.)

Note that in the above example there was no standard set as in the previous question—that is, the basis for judging whether the program was "performing in a satisfactory manner" was not specified. The evaluation design simply referred to describing status and did not imply a basis for making a judgment.

This kind of question, which describes but does not set a standard for judging, is also typical for what I have described earlier as short-term or interim outcomes (remember that term from Table O.1 in Section O?). A program might have a great number of potential interim outcomes and the evaluator and program staff would be hard-pressed to prejudge acceptability beforehand.

INTENSIVE CASE STUDY DESIGN

Let us now consider another kind of design—one that takes a different approach toward understanding outcomes. As usual, we start with the question.

> **QUESTION 6:** How well are the particular needs of children eligible for Title I, special education, and English-language instruction being met by the UTT Program academic support model?

We can certainly use any of the above evaluation designs to answer this question. However, also available to us is the *intensive case study design*. Case studies are intended to describe something in great depth. They demand a careful examination of the situation with a full understanding of the context and the use of multiple perspectives. Typically, multiple qualitative methodologies are employed in order to obtain these understandings. Thus, a case study design might use observations, interviews, focus groups, or existing data sets. You might be thinking, "Well, this sounds a lot like what we would do when planning a process-oriented evaluation." That is correct. However, qualitative methods are not reserved for answering process-focused questions alone—they can certainly be applied to the study of program outcomes as well. Likewise, quantitative data help in gaining a fuller understanding of a particular case. Causal designs and some other descriptive designs are more easily depicted in the context of measuring outcomes and provide more straightforward answers to evaluation questions. In some cases, though, the insights they yield are not as rich.

Case study designs seek to holistically comprehend the variety of stakeholder perceptions and experiences within a program. They provide more in-depth understanding and a fuller, more complete and complex picture of the situation. Qualitative instruments provide the basis for developing detailed and rich descriptions of the situation being studied.

So, for the question above, let us refer again to the four essential questions that define a design. Descriptions are less easily depicted for qualitative designs because qualitative practitioners view the design as "emerging"—continually open to modification. (I personally believe that the notion of openness and possibility of modification is [or should be] a characteristic of all descriptive designs, as the evaluator seeks to be relevant and useful.) And so, the "what," "who," "when," and "how,"—the working people of evaluation for qualitative designs—may only be stated tentatively and be subject to change as the evaluation progresses.

Rather than put these answers in a "box," I discuss them more fully in a narrative style.

Of these "working people," perhaps the most important to consider here is the "who." In most respects, all qualitative studies are case studies. They may be expansive in nature (involving many people, many sites) or smaller and less intensive. Moreover, all case studies consist of subcase studies embedded within them. If there are interviews to be conducted, *who* is to be interviewed? If there are observations, then the "who" may refer to groups (grammatically, that is a "which"). So perhaps the major question for qualitative evaluations is related to the selection of cases (or subcases) to be studied.

A FEW WORDS ABOUT SELECTION

I pause here to reiterate a few comments on the ways in which individuals or groups might be selected—a technical aside, so to speak. In making selections of individuals or groups, you must again be guided by the question along with the political/social context that surrounds the question. You want to select cases that will yield the most information related to your question. In some situations, having a great deal of heterogeneity will help to capture themes that cut across all individuals or groups. Perhaps your question deals with how particular subpopulations are engaging in the process or performing—so you would be guided by this. Perhaps it is important to provide a single case in very rich detail, and focusing on a "typical" case would be most helpful. Possibly, to be convincing to potential users of the evaluation report, it is important to focus on politically important or otherwise critical cases. Another selection procedure sometimes employed is called snowball sampling (think of a snowball rolling downhill). The idea here is to have the *most* information. Each respondent is asked to suggest others who know a lot about the topic of interest.

These are but a few suggestions related to case selection strategies. Ultimately, you must be guided in your thoughtful selection by the concern for the issue that is to be addressed and what evidence will be persuasive.

And now to return to the *who*. In the case of the UTT Program, I might recommend interviewing teachers, administrators, and parents at one of the partner schools. The teacher group would be a sample representative of the diversity of faculty in experience, education, race, and ethnicity. The number of administrators at the single school site is quite small so all of them might be interviewed. Parents might be interviewed in focus groups consisting of 5–10 parents each. Each group would be relatively homogeneous, especially having groups of parents with children in Title I, special education, or fluent in languages other

than English. Of course the students who are the "who," in this case, are students eligible for Title I, special education, and students who speak a second language at home. Another kind of "who"—although not really "who"—is the classes to be observed.

What is *what*—that is, what data are to be acquired? From teachers and administrators, we might seek information on their perceptions of the effectiveness of the program for the school's population, generally, as well as for the targeted student groups. From parents we might focus on perceptions about their child's learning experiences; for students, test scores by subgroup and by special skill area, over time; for classrooms, intense observation of one representative class per grade level to note such things as character of services provided, time allocated to different subject areas, and engagement of different kinds of children in the curriculum.

When? Interviews should take place sometime between the middle and end of the school year. Classroom observations would be made about once a week for half of the school day during the months of October, November, February, and April. Test score analysis will take place after the return of year-end test results.

How? How analyzed? Data analyses are of various types. A full description of applicable procedures for quantitative and qualitative data analysis is presented in Sections R and S, respectively, but briefly, in the case of observation data, systematic analysis of field notes will provide a portrait of what goes on in classrooms. Interview data will be carefully scrutinized in order to determine descriptions by stakeholders, in their own words, of what transpires in the program. These will be synthesized into persisting themes. Test scores will be examined to identify strengths and weaknesses of the various subgroups of students.

MIXED METHODS

Evaluators have recently given prominence to a concept called "mixed methods," which involves the use of multiple data sources, multiple ways of collecting and analyzing data, and multiple designs to answer evaluation questions of interest. The reasoning behind mixed methods is rather straightforward. It starts with the notion that there is *no single appropriate data source* and, therefore, *no single appropriate design* for most evaluations. There may be different designs related to specific evaluation questions, but not a single design for all questions. Second, the idea of mixed methods proposes that any question might require *multiple methods* of different types.

Consider the first of these. Designs should match the question, the information needs of stakeholders, and the possibilities of the program context. Since an evaluation consists of multiple questions to be answered, a program evaluation may consist of many traditionally thought-of designs, each concerned with answering a specific question. Moreover, it is possible that one basic design may be applicable to multiple questions.

Furthermore, in answering a single evaluation question, there may be a need to use several different measures to examine the same question. For example, when a concept embedded within a question is complex, or difficult to measure (as in Question 6), then multiple complementary approaches might be employed to examine the various facets of the question. In doing so, the evaluator would gain the required understanding needed to properly respond to the question.

You, as the evaluator, should be guided by our four questions (i.e., what, who, when, and how). For example, many times I have used a questionnaire to acquire a general overview of views and perceptions of a particular stakeholder group. Then, armed with insights provided by the survey about ambiguities, disagreements, or less-than-fulfilling understandings, I conducted interviews on specific topic areas. Sometimes, instead of interviews, focus groups were used. Possibly quantitative or documentary records added further insight.

By this point you might be wondering, "Well, Marv, how do I know what kinds of data are important? And what does this mean for the evaluation budget?" (I love that you are practicing active reading.) These are great questions, but there is not a single correct answer because it all depends on your context. A good starting point, however, is your stakeholders. Engage them in a discussion about what data would be most useful to them. In the process of doing so, you might present a menu of possibilities and think with them about the advantages, disadvantages, feasibility, and challenges of each type of data. As you are discussing, consider the costs of each option. It is entirely possible, for example, for stakeholders to want individual interviews done. If resources (i.e., time, personnel, and finances) are prohibitive, you might propose some combination of surveys and focus groups instead. Likewise, it might be ideal to request student performance data from the local school district, but if charges are attached, you might opt to work directly with teachers to obtain that information.

As you can see, engaging in mixed-methods evaluation can, in some ways, mean doubling or tripling the work, but what you get in return—that you do not get with simpler qualitative and quantitative designs—is depth in understanding. And because resources are finite, you will be faced with the need to *trade off* what is ideal with what is

possible. This is not a difficult message for you, who have been reading this book—it captures what I have been imparting throughout. The reason why the topic of mixed methods arises in the evaluation profession is because there are many experienced evaluators who approach the field with a particular paradigm preference that they find difficult to escape—that is, "I do experimental studies." "I am a great fan of qualitative methods." See, aren't you glad that you are new to evaluation, so that you can start out with a fresh look!

A NOTE ON CAUSALITY AND CAUSAL MECHANISMS

Recall from Section I how the evaluator attempts to seek clarity from staff and other potential users while developing the logic model about why they believe the outcomes will be achieved from a specific activity/output. That discussion is worth revisiting, in part, because these beliefs will influence how you, as the evaluator, come to help stakeholders to operationalize outcomes so that they can indeed be measured. Understand that the "program theory" that stakeholders hold to be their truth is important here.

The program theory is the staff's and primary users' beliefs about how activities ought to be sequenced and what causes the outcome. It is necessary to specifically and completely address the particular reasons why it is strongly believed that the outcome can be attained from the stipulated activities and/or outputs. When we inquire further about these assumed reasons, we are *also* questioning the "causality" issue. It is safe to presume that these conversations will be most frequent and intense at the start of the evaluation planning phase, and will certainly continue throughout the evaluation itself. However, you will reach a point where sufficient information has been gathered and understandings established for you to develop an evaluation design. It is also likely that once you reach this point, you will also be able to visually depict the program theory. Figure P.2 is one example.

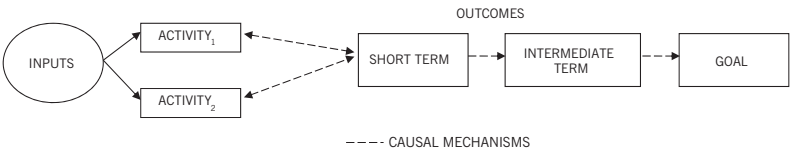

FIGURE P.2. Simplified program theory diagram.

Note that in this model, I have left off outputs but have introduced the notion of causal mechanisms. These are the justifications for why program developers believe that activities will lead to short-term outcomes, and they in turn to intermediate-term outcomes, and consequently to achieving program goals. Your job as an evaluator is to continually push for explication of the causal

mechanism and the rationale for why it leads to the next part of the theory sequence. It is possible that the connections being made are tenuous or misinformed. Measuring outcomes that have been derived from a faulty theory will be problematic for the program and the evaluator as this could potentially set both up to fail—consequences that no one wants. Thus, evaluators frequently find themselves engaging with stakeholders as *thought partners* to straighten out the underlying logic *before* any measurement or discussions about causal designs even take place. Furthermore, it is appropriate for the evaluator to suggest what is known from research and might be specifically applicable to the program context in these early conversations. Equally important, though, is the need to remember at all times that the purpose is not to create or validate research knowledge—rather, the focus is to aid in the intelligent design and implementation of the particular program. In all of these early conversations, the evaluator is trying to address the question, "What is the *presumed* causal action (or mechanism) that accounts for an activity leading to an outcome?"

SUMMARY

I have taken great latitude in the traditional descriptions of designs in order to make them more accessible. I believe that what is presented is a correct representation of traditionally discussed designs. To recap, I presented several of the more commonly used causal designs including a brief discussion of what is called the interrupted time-series design. Furthermore, I have suggested several situations where descriptive designs are applicable. Then, I commented on the case study procedure. Finally, I indicated that you might answer questions more ably by using mixed methods. Each question might require its own method and some questions might require multiple approaches.

RECAP 2—SECTION P

Evaluation Design

- Outcome Definition
- Evaluation Plan and Evaluation Design
  - Guiding principle (what, who, when, how)
- Some Causal Designs
  - Pretest–posttest comparison group
  - Posttest-only control group
  - Interrupted time-series design

- Some Descriptive Designs
  - Comparison against a specified standard
  - One-time snapshot
  - Intensive case study
- Mixed Methods
- Causality and Causal Mechanisms

➢ ***My Advice:*** Where do you go from here? You certainly may refer to the Further Reading for this section for more in-depth knowledge—and I would encourage that. However, most evaluations that you might do will be descriptive—in small-scale program evaluations you will probably not be seeking to determine causality. As a general rule, I urge you not to get caught up in design titles. At this time, I suggest that you direct your attention to thinking about design as focused on "why" (to answer questions) and then be guided by the four design questions—what, who, when, and how.

ANSWERS TO QUESTIONS 3–5

QUESTION 3: Are reading scores among partner school students improving?

What data are needed?

- Reading scores.

Who will the data be obtained from?

- Students in the UTT Program and other tutoring program classes.

When will the data be acquired?

- Five different points in time after the program started.

How will these data be analyzed?

- Comparing data after start of program with trend line.

QUESTION 4: Is a greater percentage of students performing at grade level or above in reading at the conclusion of 1 year of the program compared with the prior year?

What data are needed?

- Percentage of students performing at grade level or above in reading.

Who will the data be obtained from?

- Students at partner schools.

When will the data be acquired?

- Scores at end of program year and end of prior year.

How will these data be analyzed?

- Direct comparison of percentages.

QUESTION 5: Do students, parents, teachers, and administrators view that the program is performing in a satisfactory manner?

What data are needed?

- Attitude measure related to satisfaction with program.

Who will the data be obtained from?

- Sampling of students, parents, teachers, and administrators of partner schools.

When will the data be acquired?

- Any specified time (most likely middle or end of program year).

How will these data be analyzed?

- Summary of data by group. Examination of individual items.

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

Let us continue our discussion from the previous section, but now with an eye on program outcomes. We assume, for the sake of discussion, that Family Matters commissioned a 2-year evaluation for the purposes of understanding the RUPAS Program's effectiveness and to inform decisions about budget allocations for the next few years. We also assume that our primary users defined a "successful" program as one that can demonstrate gains in learning and skill among parent leaders and atypical levels of school readiness among children that other comparable programs do not. With these criteria in mind, how might we proceed? Perhaps we first engage relevant stakeholders in a discussion about the extent to which there is a logical flow between program activities and expected outcomes. Now, what if in the course of this discussion we learn that such a flow does not exist? Alternatively, what if we learned that program activities do in fact support attainment of outcomes? Which evaluation designs could we then draw upon to determine the RUPAS Program's success?

In the same vein, what existing data might be helpful in shedding light on this question? If you had to locate existing instruments, where might you look and which tools would be best to use? Or if you needed to develop new tools, what would you try to measure and how? As you go about this process, consider also the extent to which you, the evaluator, and program stakeholders are privileging particular evaluation designs rather than its ability to answer evaluation questions of interest. And, again, whose views are represented at this stage of the evaluation? What are they and how might they influence the evaluation's trajectory?

Further Reading

Bamberger, M., Tarsilla, M., & Hesse-Biber, S. (2016). Why so many "rigorous" evaluations fail to identify unintended consequences of development programs: How mixed methods can contribute. *Evaluation and Program Planning, 55,* 155–162.

This article is a thoughtful discussion about the ways in which mixed methods can tell a fuller story about the successes and challenges of a program than quantitative or qualitative methods do on their own.

Cook, T. (2007). Describing what is special about the role of experiments in contemporary educational research: Putting the "gold standard" rhetoric into perspective. *Journal of MultiDisciplinary Evaluation, 3*(6), 1–7. Retrieved from *http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/36.*

Tom Cook has played an important role in the development of experimental and quasi-experimental methods. This article helps to put some of the current discussion of the role of experiments in perspective.

Greene, J. C. (2005). A value-engaged approach for evaluating the Bunche–Da Vinci Learning Academy. *New Directions for Evaluation, 106*, 27–45.

Greene's article is among a collection of papers that illustrates how one evaluator chooses to evaluate a hypothetical public–private educational partnership.

Henry, G. T. (2005). In pursuit of social betterment: A proposal to evaluate the Da Vinci learning model. *New Directions for Evaluation, 106*, 47–63.

Henry's article was published alongside that of Greene's and offers a different perspective on the evaluation of the same partnership.

Mertens, D. M., & Hesse-Biber, S. (2013). Mixed methods and credibility of evidence in evaluation. *New Directions for Evaluation, 138*, 5–13.

This article provides a nice discussion of the importance of and challenges related to the use of mixed methods in evaluation.

Scriven, M. (2008). A summative evaluation of RCT methodology: An alternative approach to causal research. *Journal of MultiDisciplinary Evaluation, 5*(9), 11–24.

This is a marvelously complete discussion disputing many of the claims that are made for RCTs.

Quick Reads

1. Nicole Holthuis on Lessons Learned from Measuring Intermediary Outcomes
   http://tinyurl.com/jan4xno

2. Stacy Johnson and Cami Connell on a Mixed-Methods Approach to Data Collection
   https://tinyurl.com/m7tjkm8

3. Ann Lawthers on Triangulation Using Mixed-Methods Appeals to Diverse Stakeholder Interests
   https://tinyurl.com/kshwudj

SECTION

How Do We Manage the Evaluation?

In the previous two sections we talked about the evaluation design and recognized that it is a portion of a total evaluation plan. The evaluation plan in its totality is intended as a description of the way in which the evaluation will be conducted. But wait, might you have specified this in the contract that we discussed in Section D? That depends on the way that you and those who contracted with you viewed how the evaluation would take place. It sure is confusing. What you do now is dependent on how you started the process. Remember in Section D how you developed a proposal that in turn led to a contract? I offered three possible options that might have occurred at that time. First, general agreement was reached that you would conduct the evaluation in the manner presented in your proposal and a specific cost was agreed upon with only a slight possibility of modifying the cost or the work to be performed. A second option was an agreed-upon contract with the evaluation work specified and the possibility of making minor modifications to the study after an increased understanding of the program and stakeholders' evaluation needs had been acquired. The third possible path that the contract and the evaluation might have taken was based on a recognition that it would be important to engage in a process to better understand the anticipated operations of the program, and from that understanding develop relevant questions and an evaluation design that would be the focus of the evaluation.

This latter path is the process that we have followed in the various sections of this book. However, if either of the other paths had been followed in the evaluation, then what we discuss in Sections O, P, and this section would have been dealt with much earlier—that is, for example, if there had been a request for proposal (RFP), then the basis for your selection to be the evaluator would have been the evaluation design and the plan for fulfilling that design. Thus, issues from these sections would have been already specified.

For our purposes, we consider that the process engaged in follows the format of this book (but remember, this is not the only way that the evaluation might have transpired). And so now I present a description constituting the plan for the evaluation. In part, this description reflects upon those evaluation activities that have already been completed in preparation for developing the design. In part also, it specifies and anticipates aspects of the evaluation yet to come. By way of review, I have fully depicted this sequence in Figure Q.1 (a duplication of the Overview Chart on p. 3). In the evaluation plan, you will need to provide some general insights into how you anticipate that the future activities will take place.

EVALUATION ACTIVITIES: PAST, PRESENT, AND UPCOMING

As can be seen in Figure Q.1, the different evaluation activities take place at various stages throughout the process of conducting the evaluation. Activity 1 reflects how you get to understand yourself as an evaluator—your personal cultural context. By this I refer to your strengths and weaknesses, your background, your biases, and so on. Activity 2 addresses aspects involved in acquiring and contracting for the evaluation. Activities 3–5 are also primarily conducted in the preplanning stage of getting acquainted with its stakeholders, its context, and the program, respectively. This is indicated by the word "primary" in the preplanning stage column. Attention is also paid to these activities subsequently as indicated by checkmarks.

Activities 6–9 primarily occur during a "getting started" phase of the evaluator's work. This is indicated by the word "primary" in that column. Details related to Activities 1–9 have been presented in previous sections of this book.

Activities 10 and 11, pertaining to this section and the prior two, are devoted to writing the plan down. In Activity 10 (Sections O and P), I discussed the technical aspects of the evaluation design. I focused on the selection of process measures and on the general methodology for

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. Who Is the Evaluator?	Section C	Understanding who you are as an evaluator			
2. Contracting for the Evaluation	Section D	Primary	✓	✓	
3. Identifying Stakeholders	Section E	Primary	✓	✓	✓
4. Gaining Understanding of the Organizational/Social/Political Context	Section F	Primary	✓	✓	✓
5. Describing the Program	Section H	Primary	✓	✓	✓
6. Understanding the Program	Section I		Primary	✓	✓
7. Developing Initial Evaluation Questions	Section J		Primary	✓	✓
8. Considering Possible Instrumentation	Section K Section L Section M		Primary	✓	✓
9. Determining Evaluable Questions	Section N		Primary	✓	✓
10. Finalizing the Evaluation Plan (Design)	Section O Section P	(Primary)	(✓)	Primary	✓
11. Managing the Evaluation	Section Q	(Primary)	(✓)	Primary	✓
12. Analyzing Data	Section R Section S			✓	Primary
13. Answering Evaluation Questions	Section T			✓	Primary
14. Reporting Evaluation Results	Section U			✓	Primary
15. Helping Stakeholders to Use the Results	Section V	✓	✓	✓	Primary
Aids to getting it done properly					
Strengthening Relationships with Stakeholders	Section G	✓	✓	✓	✓
Abiding by Appropriate Evaluation Standards	Section W	✓	✓	✓	✓
Additional evaluation option					
Conducting a Cost Analysis	Section X	✓	✓	✓	Primary

FIGURE Q.1. Overview chart: Evaluation fundamentals.

considering outcome data to be collected (what), from whom the data will be collected (who), the time schedule for data collection (when), and the manner in which data will be analyzed and values attached (how). In this section, I follow up by suggesting how activities to be subsequently conducted (such as data analysis and valuing) are to be previewed within the evaluation plan. The full discussion of data analysis is presented in subsequent sections, but it is necessary to provide an indication within the evaluation plan of how that might take place. In this section, too, I deal with some of the procedures for conducting the evaluation, including agreements to be reached. Some of this might have been covered in preparing and agreeing upon a contract and a tentative budget (Section D). Now is a time to reexamine some of these issues in light of the design developed in Sections O and P. A further part of this section deals with issues related to the ongoing efficient management of the evaluation.

Activities 12–15, focusing on analyzing the data and continuing through to the activity of helping the stakeholder to use the results, are only anticipated at this stage of the plan development. The evaluator's job is to describe how each activity *might* take place. As indicated in Figure Q.1, these activities primarily occur in the "executing the plan" phase of the evaluation. The exception is Activity 15, which deals with helping stakeholders to use the results. As you have undoubtedly come to understand from previous chapters, the evaluator's role in promoting evaluation use is continual.

A further note here is helpful. I have shown in Figure Q.1 that Activities 10 and 11 might take place in the preplanning and getting started phases of the evaluation if your contract had been based on a prespecified evaluation plan at the outset—as I have discussed previously. This is shown in Figure Q.1 by providing the alternate start (in parentheses) for Sections O, P, and this section.

Two further listings in Figure Q.1 describe the aids to getting it done properly. One of these refers to the evaluation standards, which is discussed in Section W. It is important that the evaluation plan be constructed in a manner consistent with the standards of the profession. Section G, which we have already discussed, indicates the importance of maintaining positive interpersonal relationships throughout the whole evaluation.

One additional evaluation option is presented. This is an evaluation activity that might, or might not, be part of the evaluation. I refer here to the conduct of a cost analysis study. I have listed this as taking place primarily in the "executing the plan" phase, but if it indeed is to take place, it should be discussed fully in the evaluation plan.

THE WRITTEN EVALUATION MANAGEMENT PLAN

In the previous sections I discussed the evaluation design—the technical aspects of what you will do in the evaluation. These included the evaluation questions, data to be collected and from whom, and the way you anticipated that data might be analyzed to answer questions. Now it is time to add to this design the procedural aspects of the evaluation. Some of these we have already talked about, but now it is time to *write them down*. Let us examine the various issues that need to be considered and incorporated into the evaluation management plan. Start by examining the following list.

Procedures and agreements

- What is the time frame of the evaluation and the resources available?
- Yet again, we examine whether the questions and design are appropriate.
- What staff resources will be required to conduct the evaluation?
- What space will be needed?
- What equipment and materials will be needed?
- What is the schedule for reporting and the required products?
- What are the organization's administrative procedures?
- What are the responsibilities of the program and of the evaluator?
- How will differences be resolved?
- What are the standards for judging the evaluation?

Perhaps the first thing to do is to make sure that there is agreement about the *time frame* of the evaluation and the *resources* available. When does the evaluation begin? When does it end? What financial resources are available for the evaluation? In many instances, the evaluation would have begun prior to the development of the evaluation plan. The evaluator might have been hired based on a program proposal, which was subsequently subject to revision, or the evaluator might have been hired with the idea that part of the evaluation process involves the development of evaluation questions and the evaluation plan. Presumably, also, the designated amount of financial resources available should have been agreed upon—this needs to be clarified and adjusted if necessary and possible.

You have engaged the program's primary stakeholders in a process of (hopefully) constructing a logic model and (certainly) the development of evaluation questions presumed to be evaluable, and you have developed an evaluation design appropriate for acquiring answers to those questions. One further double-check is not inappropriate—do a *final review* of *questions* and the *design* selected for answering those questions. Is this what they really want? Engage the primary stakeholders in considering the shortcomings and strengths of the methods for answering the evaluation questions. It is better to talk about it now than to be criticized later. This may seem redundant, but its importance bears further repeating. Also, revisit *agreements* about the *system for stakeholders valuing* the findings and reexamine the *potential encouraged uses* of those valued findings that had been derived from scenarios you presented.

In considering the design, determine the appropriate timing for accomplishing each of the evaluation tasks entailed. Are there instruments that need to be further developed? When will this take place? When will data collection take place related to each of the questions? How much time will be required for data analysis, valuing, and writing a final report? What modifications need to take place in order for everything to be accomplished by the end of the contract year? We consider a more detailed management time line later, but first let us examine several issues to be resolved.

Now that you have an idea of potential timing issues, it is important to further consider the *evaluation staff resources* that will be needed to conduct the evaluation. You probably have already engaged staff, or at least thought about it. But the development of the time line may point out the necessity for extra staff at certain points in time. Consider whether you will need additional staff to assist in data collection. There may also be a need for particular staff capabilities, or perhaps for staff training.

What *space* will be necessary to accomplish the evaluation? Presumably, you have some office space available for staff to work comfortably on the project. This should be discussed and included in the plan. Will you need physical space at the program site?

What *equipment and materials* will be needed for the evaluation? Will there be a need for computers or equipment for administering tests or questionnaires? Will these need to be available and easily accessed? You should consider whether you might need some special technology or computer applications to perform the appropriate quantitative or qualitative analyses. Also, don't forget materials—test instruments might need to be purchased. Who will bear the cost? Think now of what other materials might be required.

You need to reexamine agreements made in the contract with respect to *reporting requirements* and *required products*. What reports are

expected and when? Is there a specific report format that is anticipated? Who has the responsibility for printing written reports? How, and to whom, will they be distributed? What are the dates when reports, of all types—written or oral—are expected?

A very important element in the contract is the program's *administrative procedures*. Many an evaluation has been unsuccessful because the evaluators did not determine ahead of time what administrative procedures would be required in order to conduct various activities within the evaluation. Whom do you report to? Who authorizes data collection? Are there administrative constraints that might cause impediments to the evaluation? Were your understandings at the time of the contract correct? Is it time for some further clarifications in the contract?

It is also extremely important to reexamine understandings about the *responsibilities of the program* as well as the *responsibilities of the evaluator*. This is not an issue that you want to have to negotiate during the process of collecting data. For example, will program staff members schedule the required interviews? Will the staff members or administrator of the program be responsible for ensuring that program participants are available during data collection periods? Generally, will the primary stakeholders and others cooperate with the evaluation and provide necessary assistance? Other matters that I have referred to earlier also pertain here. Who will print reports? Will space be provided at the site?

As the evaluation proceeds there may be a need to modify elements of the evaluation. Perhaps the data collection will become too time-consuming. Also, possibly there will be dropouts from the program that in some way impact the design. There may be differences of agreement about how or whether changes will be made. A procedure for considering the *way in which disputes will be resolved* is an important element of the evaluation plan. Admittedly, as careful as you are to be inclusive, whatever is written will not completely address the resolution of differences—but some attempt at addressing this issue is worthwhile.

When you do an evaluation, you obviously want to do a job that is considered "well done." What are the *standards* that will be employed by the program client in making *judgments* about the evaluation? I refer not to the standards for judging program outcomes, but the standards for judging you and your work. What will it take to get a pat on the back and a statement of "Job well done"? (Actually, it doesn't quite happen that way, but you get what I mean.)

The evaluation plan is an agreed-upon description of what you will do in the evaluation, what others will do, and how a successful evaluation can be achieved. In lieu of a recap of what we have discussed, see

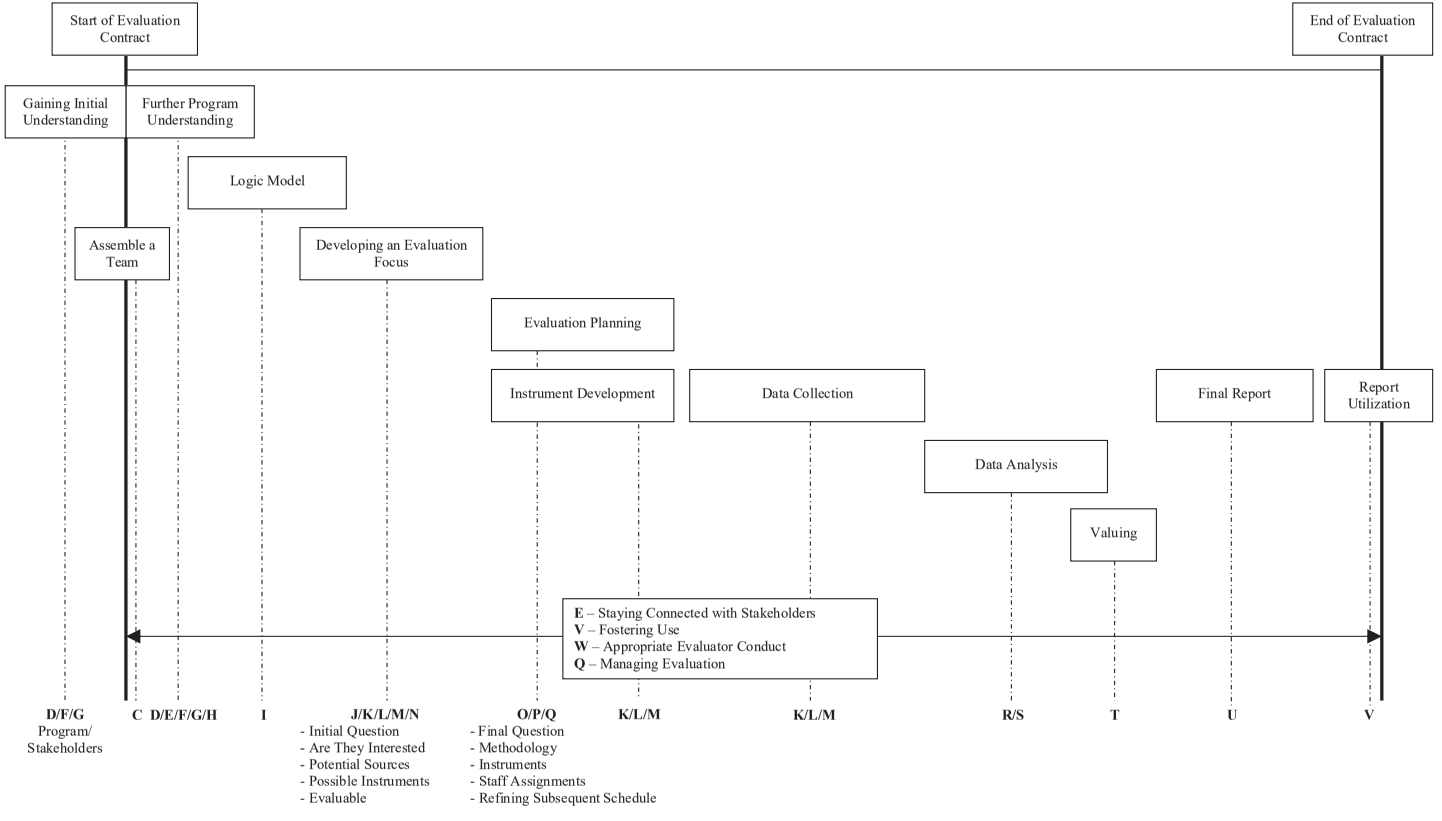

FIGURE Q.2. Evaluation management time line.

the list earlier in this section—read it and test yourself on how much detail you can remember!

OPERATIONAL MANAGEMENT

You now have an evaluation design and a plan for conducting the evaluation. There are a variety of activities associated with successfully managing the evaluation. First and foremost is getting a handle on the tasks to be completed. Then, there is a reexamination of the evaluation personnel and their fit for the tasks. Consider now what problems might arise so that you can maintain the schedule, and of course (as I constantly preach), there are relationships to be maintained. Let us consider some of these management issues—although you will certainly think of many others that need to be examined.

Evaluation Tasks and Their Sequence

The evaluator must delineate the evaluation tasks and determine how long each will take and the necessary deadlines for each. Much of this may already be specified as part of the evaluation plan deliberations discussed earlier—the sequence of tasks will, more or less, coincide with the order presented in the sections of this book. In developing the sequence of tasks, there is a need to indicate which evaluation personnel will be involved in each task. Finally, there is the matter of keeping track of all of this to see that it is proceeding as planned.

However, it is not that simple to determine when each task will take place. The sequence suggested in this book is of some assistance, but there are contingencies. Some evaluation tasks must be completed before others can commence. Why not make a list of all the tasks that need to be accomplished and then consider which ones precede others? Be especially alert to contingency activities—for example, Activity X might need to be done before Activity Y can be started—be aware of these prerequisite activities. Or, perhaps, some activities might need to be started before others but do not need to be concluded beforehand.

Now you might prepare a chart showing a time line for the project that depicts the sequence of activities that must occur. In doing so, you must consider the amount of time required for each task, the general sequence of tasks, and the contingencies. Perhaps you might construct a chart as noted in Figure Q.2.

As an example, Figure Q.2 provides an *evaluation time line* that is a guide for managing the evaluation. Notice that I have not provided dates for the accomplishment of each of the evaluation tasks. Nor have

I provided an indication of the amount of evaluation personnel time (or days) to be allocated to each task. This is intentional, since I believe that each evaluation has its very own idiosyncrasies. Some evaluations may start with an RFP and a proposal, which might correspond to what I have designated as "evaluation plan" on the time line. Furthermore, some other programs to be evaluated might already have a logic model and will not need to have evaluator assistance on that action. In other programs, primary stakeholders might adamantly refuse to engage in the logic model exercise. Some evaluations might require a vast amount of instrument development while others might use mostly existing instruments. Some evaluations might have been operating for some period of time and have already gathered substantial process data for determining program implementation. I think you get the point—every evaluation is different. Thus, I ask you to consider this general evaluation time line as simply a stimulus to your own thinking as you engage in an evaluation.

As you may have noticed, I specified several items not typically found on an evaluation time line. Two of these precede the start of the evaluation contract, one of which is "gaining initial understanding" of both the program and the stakeholders. The second, "assemble a team," begins prior to the contract, but continues thereafter. At the end of the time line, an activity—"report utilization"—appears that is listed as extending beyond the end of the evaluation contract. With respect to that task, I believe that "helping stakeholders to use the results" is an essential part of the evaluator's job at the very start of the program, throughout, and even beyond the end of the contract.

To a great extent, the general time line presented in this chart follows the sequence of sections presented within this book. I have paid heed to the natural dependencies between tasks—that is, which tasks must precede others. Furthermore, I have noted instances where a task may commence prior to the start of another task on which it is dependent, but will continue subsequently (e.g., see the "data analysis" task). Finally, I so value the importance of the evaluator maintaining appropriate relations with the stakeholder that I have designated this as a task that continues throughout the full program contract.

Note the key below the time line chart to see which sections of this book deal with each task.

Reexamining Staff Skills

Another management task associated with the time line that you may now have constructed is the determination of the skills necessary for carrying out aspects of the evaluation. You may have initially selected

staff members for accomplishing the evaluation based on what you believed to be the needs of the project. After now completing the evaluation plan, you will have gained a better understanding of the specific evaluation tasks that need to be conducted. Thus, it will be important to consider anew the staff requirements necessary for fulfilling those requirements. You must understand your staff's skills well enough so that evaluation personnel can be assigned most appropriately. In some cases, you, as an individual evaluator, or as a member of a small team of evaluators, might feel that you do not have sufficient competency to properly accomplish particular tasks. Staff training may be necessary. You should consider staff already on board to determine whether they have the appropriate skills—technical, social, and cultural. Examine accurately and forthrightly your own skills as well. You may need to reassign task assignments or perhaps there is the necessity of recruiting additional evaluation staff. Furthermore, consider the places where appropriate short-term consultants might be needed.

Maintain the Schedule

So, you have completed a schedule of activities and personnel assignments, including your own evaluation responsibilities. What's next? Accomplish it! It is important that you are attentive to *maintaining* and adhering to the *time line*. Check progress on a regular basis and if you are falling behind, do something about it. Reasons for not maintaining the schedule are many. Sometimes program staff have changed the schedule and you have not obtained proper access for data collection. Possibly, implementation difficulties have caused delays not only in the program schedule but in the completion of the evaluation tasks as well. Program-caused delays should be discussed promptly with the administrator in charge of the program (and the individual responsible for the contract, if different). Delays in communicating are not helpful. Decide jointly about what can be done and whether modifications in evaluation responsibilities are necessitated.

Other difficulties in maintaining the schedule may be attributable to failures of the evaluation team. Have there been staff losses and is there a need for more evaluation assistance? Did the plan underestimate the amount of time required to complete some evaluation tasks? Were there computer problems that slowed data analysis? The possible problems are endless. The evaluation plan and time line should have built in a little "cushion" so that minor difficulties can be handled with relative ease. Anything beyond that might require that the evaluation team provide the additional resources needed or renegotiate the budget if possible (usually not).

Communicate, Relate, Communicate

Finally, a very important management responsibility is maintaining contact with primary stakeholders and keeping an ear open to the thoughts of the larger stakeholder audience. Update and inform primary stakeholders. Be available to listen to their concerns (although, surely all of them cannot be alleviated). As I have noted throughout this book—maintaining positive relationships with stakeholders is essential, as is awareness of the cultural context.

RECAP—SECTION Q

Managing the Evaluation

- Evaluation Activities Summary
  - Completed
  - Forthcoming
- The Written Evaluation Plan
  - Writing down what has already been put in place
  - Review tasks and agreements
    - Time frame
    - Resources
    - Questions and design
    - Evaluate staff resources
    - Space
    - Equipment and materials
    - Reporting requirements
    - Administrative procedures
    - Responsibility of evaluator and program
    - Dispute resolution
    - Standards for judging the evaluation
- Operational Management
  - Tasks and time line
  - Staff skills
  - Maintain the schedule
  - Communicate

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

Managing an evaluation includes overseeing and addressing not only the evaluation's logistics but issues that could potentially impede your ability to carry out the evaluation as well. Examples include, for instance, issues related to staffing and expectations regarding project time lines and workload. Will you need an evaluation team to evaluate the RUPAS Program? How will you go about assembling one? What qualities—knowledge, skill, and experience—must each team member possess? How will each person contribute to the evaluation? What are his or her roles?

With respect to expectations, how will stakeholders' time-line concerns be addressed? Likewise, how will the team's expectations regarding workload be managed? What is the flow of communication? Have expectations been made explicit? Have any expectations changed, and how might these changes affect the previously agreed-upon evaluation contract and budget?

What are the other procedural issues that need to be addressed for this RUPAS evaluation to be considered a success?

Further Reading

Hawkins, P. (2010). Successful evaluation management: Engaging mind and spirit. *Canadian Journal of Program Evaluation, 25*(3), 27–36.

Rather than focus on the technical aspects of managing an evaluation, this article discusses the social facets of this activity. It offers a perspective that is often not discussed and is a good introductory resource.

Hobson, K. A., & Burkhardt, J. T. (2012). A lesson in carefully managing resources: A case study from an evaluation of a music education program. *Journal of MultiDisciplinary Evaluation, 8*(19), 8–14.

This article offers an example of how a process and outcome evaluation are managed, particularly in the context of a music education program with limited resources.

Stufflebeam, D. L. (2004). Evaluation design checklist. Retrieved from *www.wmich.edu/sites/default/files/attachments/u350/2014/evaldesign.pdf.*

This checklist is a helpful reference to consider when trying to wrap your head around the totality of the evaluation process.

Quick Reads

1. Kathryn Hill on Evaluation Management in Nonprofit Settings
   http://tinyurl.com/nomcg7y

2. Valerie Konar on Getting Evaluation Results through Project Management
   http://tinyurl.com/gm25ab5

SECTION

R

How Are Quantitative Data Analyzed?

Okay, I know that you're already thinking about the "S" word—statistics! But don't worry; there will be no formulas in this section. And "statistics" is not a nasty word. I provide you with some general understandings about statistics, explain a few terms, and talk about how to get started and how to think about handling quantitative data. Finally, I provide some general guidelines for what kinds of analyses are appropriate given different types of data. For the serious "heavy" stuff, you can consult Further References at the end of this section, or take a statistics course (or both).

Let's get started. By now you should know what I am about to say in the next sentence. *You must be guided by the questions* to be answered—what are you trying to find out? You need to consider how the data that you have acquired might be examined so that they will shed light on those questions. What are the outcomes that you want to examine (the *dependent variables*)? What are the activities or predictors that you believe contributed to the attainment of those outcomes (the *independent variables*)? Clearly, you should have considered these issues prior to collecting the data.

First, however, a little explanation about some statistical terms is necessary.

TYPES OF DATA

What kind of data do you have? The kind of data analysis that will be possible is determined by the kind of data that you have. So, let's talk about four different levels (or kinds) of data. The four levels that I discuss are *nominal, ordinal, interval,* and *ratio* data.

Nominal data mean that the data are named; they constitute a category. And so, a questionnaire that asks for ethnicity and provides response possibilities of African American, White, Native American, and so on is creating categorical data. Likewise, responses about political party affiliations (e.g., Democrat, Republican, Independent) and religious groups (e.g., Christian, Jewish, Muslim, given as examples of the most common religions in the United States) generate nominal data as well.

Ordinal data imply that the data follow some order—a rank order. We are seemingly surrounded by ordinal data. When asked about class ranking, for example, we think about who is at the top of the class versus the middle and the bottom. Similarly, we might think about 12th-grade students as outranking their peers in lower grade levels. When working with ordinal data, know that there is a ranking, or an order, implied.

Interval data share the notion of rank order with ordinal data, but in addition create equal intervals—that is, the distance between comparative points on a scale is the same. The distance between 75 and 80 feet is the same as the distance between 85 and 90 feet. A temperature of 90° is 5° warmer than 85°, as is 80° when compared with 75°. Not surprisingly, these data are sometimes also called numerical data.

In the same vein, *ratio data* share with interval data the notion of equidistance. In addition, though, a 0 on a ratio scale has real meaning—that is, it is the absolute absence of what you are trying to measure. Someone who does not have a pulse (i.e., zero heart beats per minute) might presumably be deceased. Similarly, someone with zero alcohol in her or his blood could be considered sober. For purposes of this analysis, I do not make the distinction between interval and ratio data.

Now that we have addressed the kinds of data that we can anticipate working with, let us talk about what we can do with them—describe, compare, and relate.

DESCRIBING WHAT'S IN YOUR DATA SET

The first step in understanding and describing data is to make them more accessible. This can be done by listing the data in order. For example, if the data are nominal (e.g., ethnic categories), then each of

TABLE R.1. Frequency Table of Nominal Data

Ethnicity	Frequency
African American	17
Asian	20
Latino	15
Native American	14
White	22
Multiethnic	35
Other	15
Decline to state	12
Total	150

TABLE R.2. Frequency Table of Interval Data

Quiz scores	Frequency
100	2
99	7
98	1
88	10
87	3
85	7
83	3
79	2
Total	35

the categories can be listed and you might tabulate how many times each category occurs in the data (see Table R.1). If the data are interval, then each of the numbers would be listed with an indication of how many times each had occurred. You would be creating what is called a *frequency table* (namely, how frequently each response was made; see Table R.2).

For interval data, it is also helpful to look at the *relative position* of scores in a frequency distribution. A person's relative position in a group may be portrayed by indicating what percentage of the people in a group has a score less than her or him (assuming that more is better). This is referred to as the *percentile rank*. Relative position might also be indicated by *deciles* (think "10") and *quartiles* (think "25"). In essence, being in the first decile means being in the first (or lowest) 10%. Likewise, being at the 75th percentile means being above 75% and below 25% of others on a (SAT, weight, height, etc.) scale.

MEASURES OF CENTRAL TENDENCY

In statistics, measures of central tendency are ways of telling what response (a category or a score) is most frequently listed, what the average response is, or what response is in the middle of the distribution. For nominal data, the best descriptor of central tendency is what is called the *mode*. Mode is simply another way of saying "style" or "typical." For example, when we say "pie à la mode" we are simply saying

pie in the prevailing fashion or style—namely, with ice cream. And so, a question about mode for nominal data is "Of the categories depicting ethnicity for participants, which one is most frequently represented?" From Table R.2, the mode would be "multiethnic."

Now what about a measure of central tendency for ordinal data? Think, for example, of an ordinal variable that consists of five categories from "strongly agree" to "strongly disagree." Clearly, we could describe the frequency distribution in terms of the mode—the most frequently mentioned category, but we are able to go beyond that. We could find the *median*. Median simply means middle (think of the median on a highway). What is the middle score? If responses were collected from 35 people on the "strongly agree" to "strongly disagree" scale, what was the 18th highest score? On the other hand, if responses were collected from 36 people, then what is the average of the 17th and 18th scores? (More on averages in the next paragraph.)

Interval data introduces additional possibilities because it creates equal intervals—these are numbers—and numbers is where statistics really gets it going. Of course, a mode or a median can be found for an array of interval data, but with interval data it is possible to calculate a *mean*. A mean is simply an average. What is the average score on the last exam in class? A mean is calculated by adding all of the scores together and then dividing by the number of participants. From Table R.2, the average quiz score in a class of 35 students would be about 89.5 points.

MEASURES OF VARIABILITY

What about *variability*? The question here is how much the scores are dispersed or spread out. Not much can be said about variability related to nominal data because they are distinct categories. On the other hand, ordinal data can be looked at in terms of its *range* (what is the lowest response and the highest response). "Oh, the scores ranged from 'strongly disagree' to 'strongly agree' " (see Table R.3), or from 79 to 100 points (see Table R.2). But this doesn't really tell you much. To get a better sense of the entire data set's distribution, you might consider looking at the *interquartile range*—the distance between the first and third quartiles. The data from Table R.2 help to determine this information. We already know that the lowest and highest scores in the table are 79 and 100 points, respectively. From our previous discussion about measures of central tendency, we can determine that the median is 88 because it is the 18th score of the 35 scores. Finding the interquartile range, by definition, calls us to divide these 35 scores into four groups. When we do so, we find that the first and third quartiles start and end

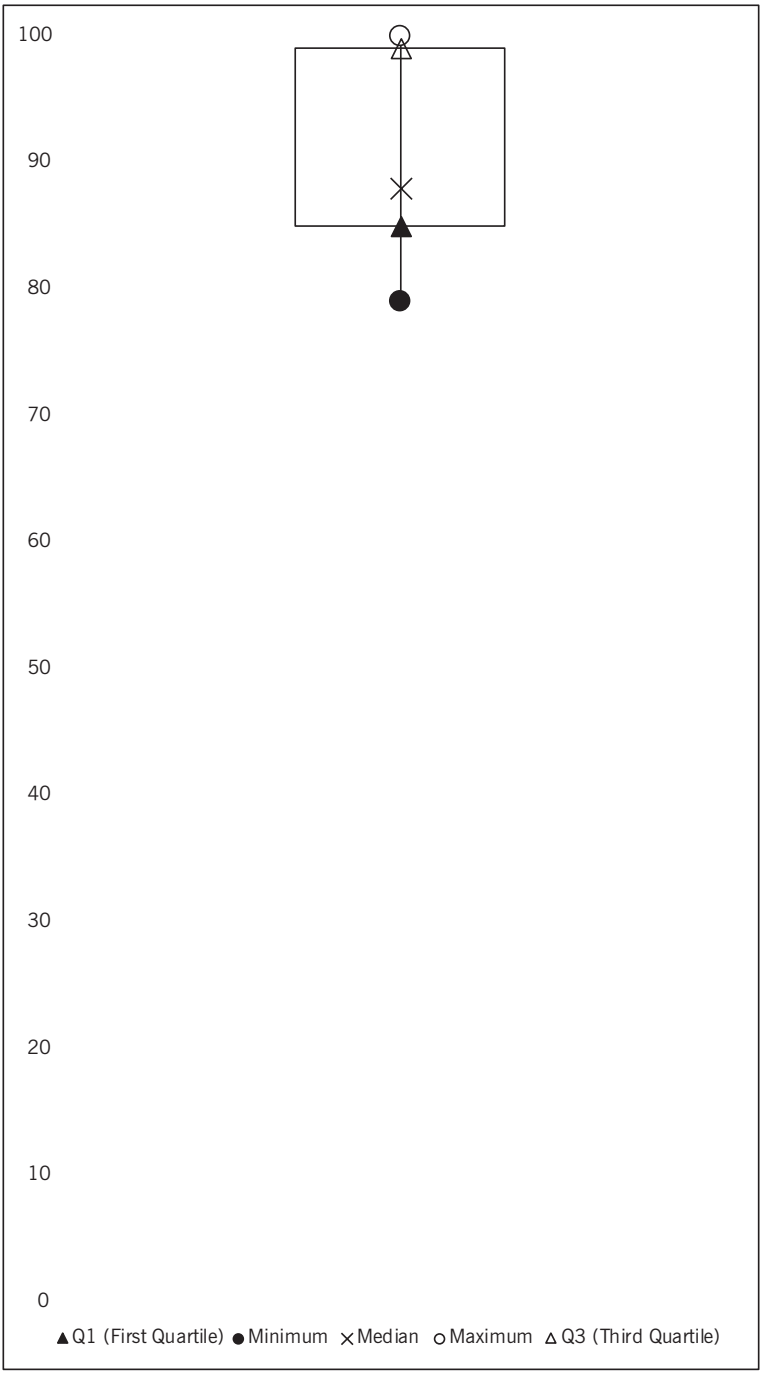

FIGURE R.1. Box plot.

at 85 and 99 points, respectively. The interquartile range, then, is the difference between these start and end points, and in this case, it is 14 points. We can visually represent this information with a box plot as shown in Figure R.1—it tells you where the bulk of the data lie.

The range of interval data does provide some meaningful information. It provides an indication of the numerical distance from the highest score to the lowest score. However, for interval data, a better indicator of variability is a statistic called the *standard deviation*. This is a number representing the spread of the scores around the mean. This statistic is important in understanding the data and subsequently calculating the probability that the results shown are real and not attained simply by chance. Given the assumption of a *normal distribution,* approximately 68% of the scores will fall within the interval encompassed by 1 standard deviation on each side of the mean, and 96% within 2 standard deviations (see Figure R.2).

These different values and ways of displaying data are called *descriptive statistics*. Again, if you have data from a total population, simply describing the data using descriptive statistics may be the appropriate way to proceed. If the data are interval, then you might want to provide the mean or median, for example, and you might want to provide the standard deviation as a descriptor of variability.

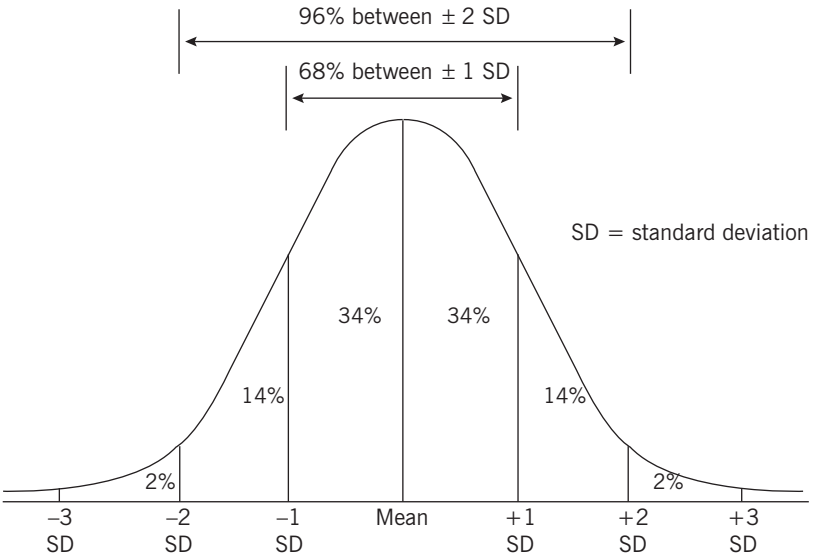

FIGURE R.2. Normal curve.

OTHER WAYS TO DESCRIBE YOUR DATA

Up to this point, we have entertained ways to summarize our understanding of the data one variable at a time. However, our data sets typically contain information about more than one dependent and independent variable. How do we begin to describe them? *Cross-tabulation tables* are especially helpful in this case. They allow you, for example, to depict the number of females who reported one of the categories from "strongly agree" to "strongly disagree" in relation to the number of males at each of those same ordinal levels (see Table R.3).

Another option is to determine the *correlation* between two variables—that is, the strength and direction of their relationship. This, of course, requires two variables that have *numbers*—interval data. If two variables have a strong, positive correlation, then one increases in value as the other increases, as shown in Figure R.3. On the other hand, if two variables are negatively correlated, then as one variable increases in value, the other decreases (see Figure R.4). Finally, as illustrated in Figure R.5, if two variables are not related in any way, then this is called a zero correlation.

You might notice that much of what we are trying to accomplish in this early stage of analysis is to understand the overall shape of the data. We have relied primarily on tables and line drawings, but other visualizations might be helpful as well. Bar charts, histograms, and scatter plots are a few examples that are often used for this purpose. Figures R.6 and R.7 illustrate the data that appear in Tables R.1 and R.2, respectively.

TABLE R.3. Cross-Tabulation Table of Agreement Responses by Gender

	Gender		
Scale	Male	Female	Total
Strongly agree	20	25	45
Agree	15	10	25
Disagree	15	15	30
Strongly disagree	50	50	100
Total	100	100	200

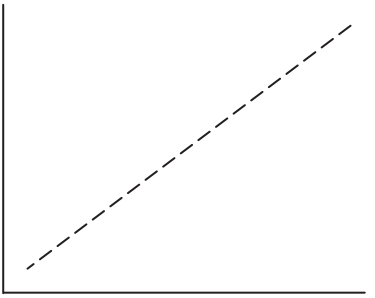

FIGURE R.3. Strong positive correlation.

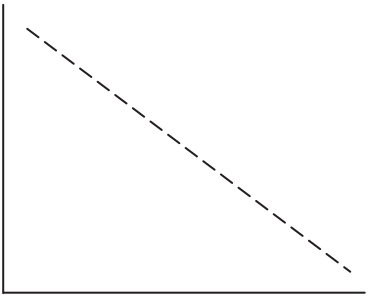

FIGURE R.4. Strong negative correlation.

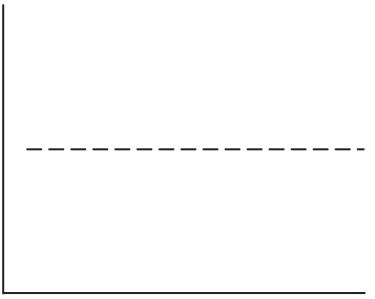

FIGURE R.5. Zero correlation.

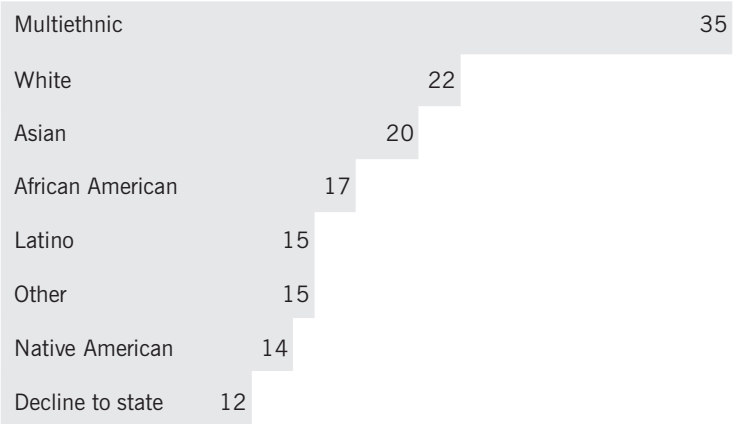

FIGURE R.6. Bar chart of ethnicity data from Table R.1.

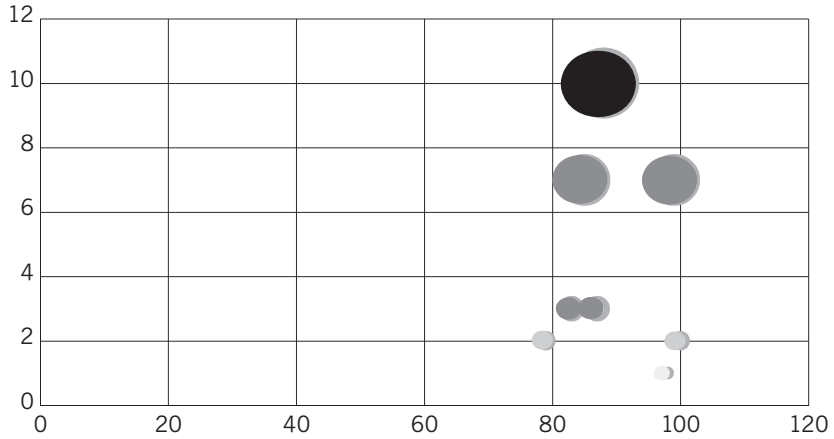

FIGURE R.7. Scatter plot of quiz scores from Table R.2.

SURPRISES IN YOUR DATA SET AND HOW TO DEAL WITH THEM

Let's stop for a moment and talk about the words *normal distribution*. A normal distribution is found when the data form a bell-shaped curve, as shown in Figure R.2. The reason why normality is important is because most statistics are based on assumptions of normality. Remember the undergraduate courses where the professor "graded on the curve"—so many percent were "A," so many percent were "B," and so on? Note that in a normal distribution, the mean, median, and mode are the same. Non-normal distributions typically are caused by the presence of *outliers* in the data set. These are data points that stand too far off from where all the others are clumped. Leaving outliers in the database in your analysis might lead to results that are *skewed*—results that appear to be higher or lower than what is actually present. For example, in Figure R.8, where the distribution is positively skewed—that is, skewed to the right—the mode is less than the median and the mean. On the other hand, in Figure R.9, where the distribution is negatively skewed (or skewed to the left), the most typical score (mode) is greater than the middle score (median) and the average score (mean). These results might be caused by outliers—scores that are vastly different from the rest of the data. (One potentially easy way to remember the direction of the skew is to ask yourself, "Where is the tail—is it on the left or right side of the curve?")

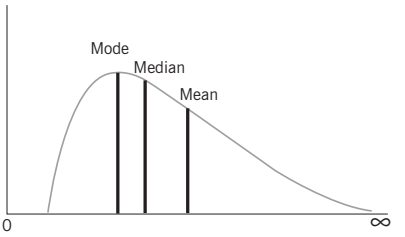

FIGURE R.8. Positive skew.

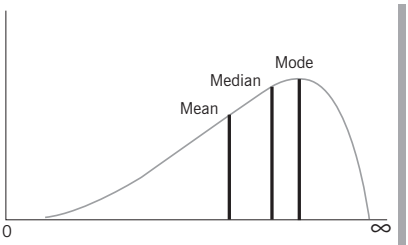

FIGURE R.9. Negative skew.

Understanding the distribution of the data, including whether data are normally distributed, is essential. If it is a *small data set* (one that you feel that you can handle manually), then you can look at the responses. Are there some data that clearly don't make sense? If there are "outliers," their presence can lead to a non-normal distribution. You might want to go back and check on whether the outliers are typographical mistakes. Suppose that your program participants (the group of people from whom you collected the data) were high school students and you see a data item that lists an age of 10, or an age of 32—clearly these are incorrect data. Are there other responses that simply don't make sense? If you can bring clarity to a particular data response, then do so—otherwise, you may need to delete that response.

Another kind of problem is determining how to deal with *missing data*—that is, perhaps a respondent failed to answer one question. Researchers suggest several approaches to handling this. One of them is to leave it be and work with the data that you have—both complete and incomplete. Another is to assign the average response, across all subjects, to that location—that is, to impute data. One other approach is not to consider the response at all, meaning, work with complete data only.

These activities are referred to as "cleaning the data." You are simply getting the data into a form that is more workable. Are there surprises in what you have observed so far? Are the surprises sufficient to lead you to question the data? If so, see what else you can do to clarify the data on hand.

If it is a *large data set*, then instead of dealing with the data manually, you will want to enter the data into a computer database for ease of subsequent handling. There are a number of options that you might then use. *Microsoft Excel*, for example, provides a way to handle most of the simple statistics of data manipulations. Specialized statistical programs are needed for more complicated analyses (e.g., SPSS, STATA, R).

First, you will want to determine disparities in the data. You might start by looking at how people typically answered the questions you asked. You will want to calculate the mean and standard deviation in order to determine whether the data are normally distributed. You will check outliers, as in the earlier discussion. Before removing those outlier data points, you want to go back and check whether any were actually typographical mistakes.

A NOTE ON POPULATION AND SAMPLE STATISTICS

Getting a good handle on what the data looks like early on is important in determining how they can be further analyzed. Being able to compare groups of participants and finding relationships between variables in the data set are some examples of such analyses. So we next have a conversation about statistical techniques and statistical significance. Before doing so, however, it is important that we consider whether the data are acquired from the total population or from a sample of it.

If the *total population* (all participants) constitutes the data set and there is a sufficiently large number of participants (on the order of several hundred people or more), then the statistical techniques that we address in the remainder of this section will by and large be appropriate. More often than not, however, we do not have access to the entire population so we fall back on *samples* (smaller groups of participants) that we hope are reflective of the population. Then, we use sample statistics (or *inferential statistics*) to estimate what might be true of the larger group. Stated differently, we are often in the position of trying to arrive at results that we hope will generalize from our humble participant samples to the greater population.

SAMPLING

Even if you are not seeking to generalize your findings to other contexts, as would be the case in most formative evaluations, it might still be appropriate to use inferential statistics. Suppose, for example, that you were unable to give a test or a survey to everyone from the larger population at your program site. Instead, you used a sampling procedure (remember this from Section M?) to obtain data from a smaller group of people who you think represents the larger group. You might determine, for instance, that you would like to collect data from these participants. Who might you consider a "participant"—program alumni, currently enrolled individuals, individuals who have participated in the program for 2 years or more? Defining the pool of people whose data you would analyze (i.e., establishing a *sampling frame*) is one in a series of critical steps toward arriving at trustworthy data (and, of course, conclusions about the program).

APPROPRIATE STATISTICAL TECHNIQUES

Not surprisingly, the statistical methods appropriate for further analysis of data depend on several factors—the questions that you are trying

to answer and the level of data that you are analyzing. What do I mean by this? Each of the combinations of nominal/ordinal/interval variables that you are interested in examining will trigger, in combination with your evaluation questions, particular statistical techniques that might be appropriate. Moreover, whether one or more independent and dependent variables are to be included will lead to different statistical techniques that are to be used.

As always, we start with the *questions* and there are usually two kinds—one dealing with the nature of the comparisons that you seek to make and the other with the relationships that you aim to identify. Imagine that you are doing an evaluation—Table R.4 shows some questions that you might need to answer. Now I want you to examine each of the questions in Table R.4 thoroughly before reading what else I have to say about each question. (*Stop now* and just read each question—the first column only.) Furthermore, it is important before you proceed that you fully understand the distinctions between nominal, ordinal, and interval data. If you need to review again, do so to refresh your memory.

Let's take a look at Question 1 in Table R.4. What are the independent and dependent variables? We want to know whether there is a difference between male and female program participants. This is the nominal, independent variable. The question suggests that we want to know whether there is a difference between extent of participation—whether they participated or not—so this is a nominal, independent variable. To analyze the data at this level, we would have to construct a cross-tabulation table and conduct a chi-square test. This statistical test and others mentioned are found in the statistics texts like those in the Further Reading for this section.

What about Question 2? Our variables are "extent of participation" and "level of attainment," and both are measured on an ordinal scale, but which affects the other? Given the way the question is worded, we clearly want to know the effect of participation on attainment—the former is the independent variable while the latter is the dependent variable. To analyze this kind of data, we would need to calculate Goodman and Kruskal's gamma.

Let's do one more—Question 3. Again, given the question, what are the independent and dependent variables? It looks as though we want to compare achievement scores of participants who are in the program being evaluated with those who are not. So, achievement scores are the interval, dependent variable. The nominal, independent variable refers to participants who are in each evaluated program.

Now you try to examine the remaining questions by looking at Table R.4. Notice in Question 4 that "different cultural backgrounds" is a nominal variable and that there are multiple groups. In Question 5,

TABLE R.4. Examples of Appropriate Statistical Techniques

Question	Question type	Independent variable(s)	Dependent variable	Statistical test
1. Do male and female participants in the program being evaluated differ in their participation in after-school programs (participator or not)?	Comparative	Nominal (1 variable, 2 groups)	Nominal (1 variable, 2 groups)	Chi-square
2. How does extent of program participation (high, middle, or low) compare with academic attainment (high, middle, or low)?	Comparative	Ordinal (1 variable, 3 levels)	Ordinal (1 variable, 3 levels)	Goodman and Kruskal's gamma
3. Do participants in the program being evaluated and the comparison program differ in their achievement scores (numbers)?	Comparative	Nominal (1 variable, 2 groups)	Interval	*t* test
4. Do participants of different cultural backgrounds differ in their achievement scores?	Comparative	Nominal (1 variable, multiple groups)	Interval	Analysis of variance (ANOVA)
5. Do high scores on the attitude survey (numbers) predict high scores on the knowledge test?	Relational	Interval	Interval	Linear regression
6. How are pretest score, age, and number of sessions attended related to end-of-program achievement scores?	Relational	Interval (more than 1)	Interval	Multiple linear regression

both the independent and dependent variables are interval data. And finally, in Question 6, observe that there are multiple independent variables and that each qualifies as interval data. Note that unlike the first four questions that we worked through, the emphasis in Questions 5 and 6 is not on determining whether there are differences between groups—that is, we are not trying to arrive at a yes/no answer. Rather, techniques such as regression allow us to understand how degrees of change in one variable affects another. Thus, the answers that we arrive at deal with issues of "how much" and are more nuanced than simply yes or no.

I want you to understand that in this section we reviewed only a few examples of appropriate statistical techniques. Other statistical tests mentioned in the tables are linear regression, where independent and dependent variables are both interval and multiple linear regression, where there is more than one independent variable. Some researchers might employ different tests to answer the questions above. To assist you, there are several computer programs available for performing the various statistical analyses, including the Statistical Package for the Social Sciences (SPSS), Stata, SAS, and R.

Now, you might be wondering why we are talking about data analysis, particularly if you are doing formative evaluations. Isn't this discussion better suited for summative evaluations? This may initially appear to be the case, but actually, the kinds of data and data analysis techniques that I have mentioned are often used in summary formative evaluations as well. Evaluators need statistics to get a firm sense of what is happening in programs. Evaluators need statistics to be confident that differences are not chance occurrences.

My Warning: You clearly should not consider yourself statistically capable based on this section. There are, for instance, other statistical techniques that are more complex, such as MANOVA, MANCOVA, path analysis, and hierarchical linear models (HLMs). Many of the programs that I have worked with tended to be relatively small in terms of the number of participants enrolled. Thus, these higher-order statistical techniques were often not needed. This suggests that if simple statistics are helpful and easily understood by stakeholders, then perhaps they are sufficient. If, however, you find yourself in a position of working with much larger programs and require more sophisticated tests, then it is best to consider getting statistical assistance.

RECAP—SECTION R

Analysis of Quantitative Data

- Types of Data
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Describing What's in Your Data Set
  - Frequency distribution
  - Percentile ranks (e.g., deciles and quartiles)
  - Measures of central tendency
    - Mode
    - Median
    - Mean
  - Measures of variability
    - Range
    - Interquartile range
    - Standard deviation
  - Other descriptive measures
    - Cross-tabulation tables
    - Correlations
    - Visualizations
- Further Issues
  - Normality
  - Outliers
  - Skewness
  - Missing data
- Population and Sample Statistics
  - Population versus sample
  - Sampling
- Appropriate Statistical Techniques
  - Comparative analyses (chi-square, Goodman and Kruskal's gamma, *t* test, ANOVA)
  - Relational analyses (linear regression, multiple regression)

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

Consider, now, the hypothetical evaluation questions you addressed in the RUPAS evaluation from Sections O and P. Some, *but not all,* of these were questions that might have been answered using quantitative data (however, you might have decided to employ qualitative methods—if that's the case, skip ahead to Section S).

Let's assume that you collected output data (e.g., number of meetings each Parent Leader participated in, number of hours of reading each child completed) as part of a process evaluation focused on fidelity of implementation. What are some techniques described in this section that you can use to familiarize yourself with this data?

Now, what if you used tests to collect data on program outcomes (e.g., Parent Leaders' knowledge of how to support their children's reading and children's school readiness levels)? How could you begin to describe the data that you have on hand? What statistical tests could you use and what do the results of each tell you? How would your decision to run these tests change if you had data on 10, 100, or 1,000 parents? Likewise, what if you had data on 10, 100, or 1,000 students?

In addition to the analytic issues described in this section, we encourage you to consider issues related to bias and statistical tools for analysis. Specifically, who is engaging in data analysis—you, another member of the evaluation team, a stakeholder, or one of their staff members? Would engaging stakeholders in data analysis influence the evaluation—if so, in what ways? For instance, how might preference for or fear of numbers affect which quantitative data are collected, the manner in which they are collected and analyzed, and subsequently interpreted? How might you account for these biases in the analysis phase of the evaluation?

Additionally, how complex are the analyses that will be conducted? Are you able to arrive at results using a widely available tool such as Microsoft Excel as you would with a license-based package such as SPSS? How does use of tools such as these affect the evaluation budget in particular and the evaluation in whole?

Resources

Microsoft Excel
office.microsoft.com/excel

R
www.r-project.org

SAS
www.sas.com

Stata
www.stata.com

Statistical Package for the Social Sciences (SPSS)
www.spss.com

Further Reading

Franke, T. M., Ho, T., & Christie, C. A. (2012). The chi-square test: Often used and more often misinterpreted. *American Journal of Evaluation, 33*(3), 448–458.

A thoughtful discussion of the chi-square statistic—how it should be understood and used—is offered in this article.

Kellow, J. T. (1998). Beyond statistical significant tests: The importance of using other estimates of treatment effects to interpret results evaluation. *American Journal of Evaluation, 19*(1), 123–134.

This article pushes readers to go beyond statistical tests as a means of understanding evaluation findings. It reaffirms the importance of weighting statistical and practical significance of evaluation results.

May, H. (2004). Making statistics more meaningful for policy research and program evaluation. *American Journal of Evaluation, 25*(4), 525–540.

This is a somewhat technical paper on how to communicate statistical findings in ways that are more understandable. I include it in this section because, in my view, statistical analyses that are to be conducted should be guided by whether they can be described in a form that is useful to those who will receive the evaluation report.

Newcomer, K. E., & Conger, D. (2015). Using statistics in evaluation. In J. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (4th ed., pp. 596–635). San Francisco: Jossey-Bass.

This is a nice overview of procedures for selecting appropriate statistical techniques. It is a further expansion of what is discussed in this section.

Quick Reads

1. Spectra Myers on Simplifying Data Analysis
   http://tinyurl.com/glboea7

2. Marina Byrd on Skills Graduate Students Should Develop before Entering the Evaluation Field
   http://tinyurl.com/jduern9

SECTION

S

How Are Qualitative Data Analyzed?

In Section R, I discussed how to analyze numerical data. Insights, however, can come from narrative information as well. In this section, I discuss how to understand qualitative data. To begin, as in the analysis of quantitative data, we consider the evaluation question that needs to be answered. Are you clear about what the question is and what data you will draw from to answer it? Do you have enough data from interviews, focus groups, observations, and the program to adequately address the question? Do you still have some issues that you are quite unclear about and feel that more information would be helpful? If time permits, attempt to gather such data. Then, convert your interview and focus group recordings as well as field notes into *transcripts*—intelligible, full-text documents—that can be read and understood by anyone. Once done, you can commence analysis—an iterative process that requires transforming the collected data into manageable form (coding), becoming familiar with what is in your data set (indexing and memoing), describing your data (finding patterns), and arriving at results. Let us begin looking at each of these.

CODING

You have before you a massive amount of information—perhaps, overwhelming. What to do? You are trying to find patterns in the data that will lead to understanding and interpretation related to the evalua-

tion question or issue. Stay focused. The first step is getting the data into a manageable form—*reducing data*. This is a process that involves dividing all of the information that you've collected into neat, coherent chunks and tagging each with a word or a short phrase that describes the main idea contained within them. Organizing and labeling segments of text in this manner is referred to as *coding.* The tags or labels applied to transcripts that you amassed are called *codes*. Codes are used as needed so that contents of the chunks that they are associated with can be identified, located, and further analyzed.

How do you begin developing codes? There are several acceptable approaches. I describe two—that of the *deductive* and *inductive* varieties. Deductive coding calls for the use of preestablished and predefined sets of labels that are appropriate given the evaluation question. They might be informed by previously conducted evaluations, published research literature, or even stakeholders' need for information. This, as you are aware by now, is often reflected in the questions that they need answered.

While deductive codes are established a priori, codes may emerge inductively as well. You want to start by repeatedly reviewing your notes. In doing so, try to understand the views and thoughts of those within the program. How do *they* view their world? In your observations or interviews, it may have become clear to you that program participants (the insiders) seem to have their own typology—that is, their own way of categorizing their experiences. *Respect the insider perspective.* Try to capture the meaning of the labels that insiders use to describe people, human events, and interactions by using their words as codes. This is perhaps the most challenging part of inductive coding, mostly because there are no guidelines to follow, no previous models to use. Still, it is feasible and useful in ensuring that participants' voices are represented in the process.

Note that inductive coding—when used in combination with deductive coding (as they often are)—positions you to account for more information in your transcripts. Deductive codes are never comprehensive; there is always something that the original code developer did not consider.

The process of applying codes itself, whether developed in an emergent or a priori fashion, requires you to look at what information is actually present and try to create labels for each important piece. Read through all of the transcripts and make some comments to yourself—indicate the names or categories in the margin. What kinds of labels or categories appear to be present and repeating across documents? Perhaps then read through your transcripts again to identify additional codes that might have been missed on a prior reading. Were these new

codes? Test them against your prior interpretations. Perhaps they suggest a variation in a previous code—sometimes an expansion of the code, sometimes breaking it into multiple codes. You should read and reread transcripts to refine the coding structure and to settle on patterns emerging from the data. This code refinement process is one of the ways that the evaluator ensures rigor.

What might this coding process look like? Let's say, for instance, that you have been charged with learning about how tutors' training contributed to their ability to teach and you had interviewed a few tutors about their experiences. You might have asked about their perceptions of how the training was structured, the materials that were used, and how they use what they learned during training in their classrooms. These three broad areas could be used as *parent codes* in the initial round of data reduction that you conduct. You will need to continue distilling the data and can do so by repeating the coding process and expanding each parent code to include *child codes*. Likewise, each child code could also be expanded to include *nodes*. Table S.1 offers one example of how this is done. Consider the nature of tutor training, for example, which can be characterized along several different dimensions. Table S.1 suggests that the training can be described in terms of its *structure,* the *materials* relied upon, and the way in which tutors *use* what was learned. These are parent codes and each can be further expanded upon. For example, as a parent code, "structure" could be described in terms of training duration and training format. These are but two possible child codes; there could be others. Likewise, each child code can be characterized in more detail. Table S.1 suggests that "duration" might include descriptive nodes such as "too long," "too short," or "just right"; whereas "format" could be thought of as large group or a small group. The coder certainly has creative freedom when developing codes. The challenge, however, is coming up with codes that are germane to the data.

Developing a coding scheme in this manner—where each new code is more granular than the next—is a part of deductive coding. Inductive coding, on the other hand, involves going about this process by identifying parent codes from nodes and child codes. It is important to note that coding is not a linear process—rather, it is an iterative one. You will see this principle underscored throughout this section.

A NOTE ABOUT CODING WITH SOFTWARE

There are computer programs that assist in the coding process. Three computer programs that are especially helpful for analyzing qualitative data are called "NVivo," "Atlas.ti," and "dedoose." However, one problem in using software programs for qualitative analysis is that there can be a reliance on "auto

coding" or "keyword coding"—that is, coding by searching for specific words alone. This is very dangerous because it assumes that specific constructs are addressed throughout your transcripts in the exact same way. Software programs should *never* be used as a substitute for carefully reading and reflecting on transcribed data. However, qualitative software programs are helpful when conducting particularly lengthy studies, studies that involve many data sources, and longitudinal studies because the data can be housed in one convenient place. *Software programs are great tools to help you to stay organized, but they will not do the work for you. When possible*, I personally prefer not to use computer programs to code my transcripts. I like to "feel" the data. I like to personally "touch" each piece of the data and feel involved in getting acquainted with them. I prefer this to the anonymity of a computer doing its work—for me, the risk for inaccurate coding or misinterpretation of context is simply too high.

TABLE S.1. Possible Codes and Coding Structure of a Training Evaluation

Codes	Code structure
Structure	Parent Code 1
Duration	Child Code 1
Too long	Node 1.1.1
Too short	Node 1.1.2
Just right	Node 1.1.3
Format	Child Code 2
Large group	Node 1.2.1
Small group	Node 1.2.2
Materials	Parent Code 2
Lectures	Child Code 1
Morning	Node 2.1.1
Afternoon	Node 2.1.2
Readings	Child Code 2
Books	Node 2.2.1
Articles	Node 2.2.2
Use	Parent Code 2
Classroom management	Child Code 1
Lesson planning	Child Code 2

INDEXING

Coding is a substantial component of qualitative analysis. It is the beginning of the next phase of the process—indexing, which is an approach to organize and prepare codes and coded transcripts for additional analyses. An *index* of codes that were used (like an index found in books) tells you where to find different concepts and relevant quotes within your transcripts. It also gives a preliminary sense of *code density*—that is, how often some codes appear and where relative to other codes. The first step in preparing an index is developing an inventory. To do so, create a table that accounts for all of the data sources that were coded along with the parent codes, child codes, and nodes that were used for each. Include, as well, the relevant line numbers for ease of location.

Table S.2 provides an example of one way in which you might accomplish this task. We note among the nodes ("too long," "too short," "just right"), for instance, where in each set of interview, focus group, and observation transcripts tutors made comments about the training's duration—down to the line numbers. The child code ("duration," in this case) reflects on a broader level what data sets were coded and relevant. We leave the space next to the parent code—"structure"—blank for notes about possible conclusions that we can draw based on the distribution of nodes and child codes (more on this in Finding Patterns, below). While this is one approach that you can take to develop an index, it is not the only way. Do what makes the most sense for you and will allow you to trace your steps later on if needed.

TABLE S.2. Sample Index of Codes

Codes	Data source
Structure	
Duration	I1, I2, I3, I4, FG1, FG2, O1, O2
Too long	I1.L4–16, I1.L33–46
Too short	I3.L117–122, I4.L15–20 FG1.L35–44, FG1.L47–51 O1.L56–66, O1.L88–95, O2.L101–116, O2.L155–160, O2.L201–300
Just right	I2.L23–25, I2.L33–37

Note. I, interview; FG, focus group; O, observation notes; L, line numbers.

MEMOING

As you can probably already tell, coding and indexing are important steps to take when analyzing qualitative data. If you have been coding and indexing notes for your own project while reading this section, you might notice that you have been met with inclinations to draw conclusions about the transcripts that you have been reviewing, the codes that you have been using, and the preliminary code distributions that have surfaced in your index. Heed my advice—resist the urge to be wedded to these inclinations—rather, document and reflect on them. Write them down—perhaps in the form of journal entries. The collection of these reflective documents is called *analytic memos*. They contain your initial thoughts and hunches, as well as reflections about potential biases that you might have about what you are possibly seeing in your data, but do not yet fully understand or for which you do not have sufficient evidence.

While I discourage blind reliance on software, one nice aspect of using it is the ability to name and link memos to notes, codes, and segments of coded text. These memos can in turn be expanded upon on an as-needed basis while you continue to review, code, and index notes. Such programs will also help you to develop indexes rather efficiently. The thoughtful coding that you would have completed will be tracked in the program itself, which will allow you to view, download, and export them for further consideration relatively easily.

FINDING PATTERNS

You might begin to notice that it is particularly *difficult to separate the tasks* involved in analysis of qualitative data. The act of coding, indexing, and memoing each begets development of additional codes as well as the collapsing or redefinition of them. Each activity will also urge you to do more of or to redo the other because you arrive at new insights and hypotheses along the way. This is an inherent, and oftentimes frustrating, part of the process. It is also precisely what I meant earlier when I noted that qualitative analysis is iterative in nature. Analysis of data is an ongoing activity. At each stage, the researcher reexamines data and searches for meaning. At some point, however, we have to stop wading in data, stop coding, stop indexing, stop memoing, and try to arrive at findings. You might be thinking, "This seems endless. When do I stop? When *can* I stop?" The short answer is when you begin to notice that the same kinds of responses are appearing again and again in your data set. You can stop when you have not been able to find any instances that counter the "trend" that you are seeing—that

is, when you have reached the point of *data saturation*. How can we tell if we have reached that point?

Looking carefully at the exact nature of these codes and their interrelationships—that is, finding patterns—is a major part of the analytic process. In essence, the approach is a combination of summarizing codes and examining similarities and differences *within* codes. You will need to carefully "look inside" each node, each child code, and the parent code to understand what exactly has been identified as important. Let's take a look at Table S.2 again. Information in the table tells us that there are many instances of informants noting that the training was too short. This code about training duration appeared more frequently than others. Ask yourself, "What does this mean?" To answer this question, you have to "dig deeper"—read and reread what was coded. Note what you have observed and what your informants have said that is relevant to the code and the evaluation question. Can you summarize these segments of text into a sentence or two? This short summary is a *preliminary theme*—preliminary because you have not yet carried out this process with other codes. And so, to arrive at other preliminary themes, repeat the process.

You will also want to look at similarities and differences *between* codes. In doing this, you will begin to develop a sense of whether and how codes overlap. You will also begin to see how often codes appear in tandem with each other (or *co-occur*). You might notice, for instance, that "just right" and "small group" codes (with respect to training duration and format, respectively) co-occur relatively often. "Too long" and "large group" codes might overlap as well. What do you make of this? After looking inside each set of coded text, you might notice that there were overwhelmingly positive perceptions about training duration among those who participated in smaller sessions whereas those who were in large groups did not feel the same way. These preliminary themes and co-occurrences are patterns in your data set. Going through the process of finding them is synonymous with looking for relationships among codes.

Once you have discovered these preliminary themes or co-occurrences, what are you to do with them? How do they help you to arrive at findings? You need to ask yourself, "Are the cross-cutting themes highlighting any particular relationships?" If there aren't any consistent patterns, then that itself is a finding and you should appropriately describe the inconsistencies as well. Based on these descriptions, you can begin to "dig" a bit more. Compare relationships within and between groups of themes. For instance, consider whether the tentative patterns already noted hold across training topics. This will add depth to your understanding of the data.

At this point, you may be wondering what should be done if you begin to notice conflicting information across themes. First, review your data and ask yourself whether they are accurate. Does the data really reflect what happened in the field? Then, consider recoding the data, but not in an effort to change the results—rather, only to help you make better sense of it. If, after recoding, you still come up with conflicting information, then you have the option to go back into the field to collect more data or to describe and report the inconsistencies. While collecting more data will be costly in terms of time and other resources, it may be necessary in order to obtain accurate information and to substantiate your preliminary findings. Again, qualitative analysis is an iterative process—you look, and then you look again.

A NOTE ABOUT VISUALIZING QUALITATIVE DATA WITH SOFTWARE

Clearly, as "old-fashioned" as I am concerning how one might go about analyzing qualitative data, I am not entirely against doing things differently. Identifying patterns and relationships between and among codes is extremely challenging work—this is where software can make your life easier. Along with helping you to stay organized, programs such as the three that I highlighted previously are helpful once you have personally coded and indexed the data because they also tend to have built-in visualization tools that will help you to "see" relationships among codes. All of the programs that appear in Resources at the end of this section allow you to, for instance, develop cross-tabulation tables that reflect the co-occurrence of codes. Some of them can also create a word cloud—a graphical representation—of keywords that appear in your data set. One could, for example, analyze transcripts and derive patterns related to frequencies of codes used or other keywords in the notes relative to one another. Of course, such visuals are helpful as *preliminary* steps toward understanding your data, but nothing beats the old-fashioned approach, though. Read, reread, and read some more.

TESTING THE VALIDITY OF THE ANALYSIS

Once you believe that you have arrived at an appropriate understanding of the themes, it is important to further test the validity of your assertions. Engaging in a kind of validation process helps to strengthen your analysis and its findings. One method of testing whether what you have produced makes sense is to be a "naysayer." Say to yourself, "What if what I found is *not* true?" Consider *competing explanations*. How might the data have been organized differently? And if you do so, what are

the alternative or competing explanations that would then have been derived? Consider a new scenario: Suppose I described the themes in this way. If I did, would the data that I have fit those themes better? Discounting the adequacy of alternatives helps to validate your findings.

Further, consider the *outliers*—negative cases that appear to lead in a different direction than your findings. In Section R, we discussed quantitative outliers. Here we are considering qualitative outliers—data that differed substantially from the other qualitative information. For example, there could be opinions of people who completely disagree with the rest of the group, and they seem to be alone in their positions. How much credence do they have? Say to yourself, "If I read the field notes again with the negative cases in mind, do I gain a new perspective?" Including and considering the handful of outliers gives you more information that could potentially explain the trends that you are seeing. Doing so gives you a more accurate understanding of people's perspectives and experiences while testing the assumptions you have made.

Another means of testing the validity of your analysis is by *triangulation*—that is, a method of checking the veracity of your results by using more than one piece of information. Information for triangulation can come from a number of different places—study data sources (or methods of collecting data), theoretical perspectives, study designs, or investigators. Triangulation of data really means looking at the information in multiple ways. That is not easy to do when focused on qualitative data. Findings from observations may differ from those generated by interviews—we talked about this concept earlier—but the differences may be attributable to a different focus for each of these investigations. Another consideration: Were there differences in the perceptions of multiple evaluators or data collectors? Did the data from observations or interviews change over time? It is helpful to try to bring together multiple data sources to increase the validity of your qualitative findings. Indeed, triangulation is a hallmark of qualitative analysis because statistical testing is impossible. The claims that you make will have greater validity when they are based on consistent patterns across multiple sources of data.

Another kind of validation procedure involves what researchers call *member checking*. This involves talking with informants (e.g., those who had been interviewed) to determine whether your notes, and otherwise recorded data, were appropriately understood and interpreted. Did you properly capture their thoughts? Are your interpretations reasonable?

You, as the evaluator, could also try out brief preliminary data summaries and possible themes with selected larger groups of stake-

holders to obtain their input. I have always believed that it is helpful to gain stakeholder input on the validity of an evaluation. I talk more about that in both the valuing section of this book (Section T) and the reporting section (Section U). The insight provided by sharing initial data and interpretations may prove to be helpful in validating or potentially making modifications in the findings. However, be careful. The selection of informants might be introducing a new source of bias to the data and, moreover, it could be politically disruptive.

A final validity test relates to *you*. In qualitative evaluation, you, as the evaluator, are as much the instrument as the interview or the observation protocol. Consider your biases and your points of view. Reexamine your notes and, in particular, your evaluator memos. Are there ways in which these views—your views—might have had an undue influence on the indicated categories? Pay heed to, and be warned by, the old adage, "If you are a hammer, everything looks like a nail."

RECAP—SECTION S

Qualitative Data Analysis

- Coding
  - Deductive coding
  - Inductive coding
- Indexing
  - Inventory
  - Code density
- Analytic Memos
- Finding Patterns
  - Within- and between-code comparisons
  - Themes
  - Data saturation
- Testing the Validity of the Analyses
  - Competing explanations
  - Triangulation
  - Member checking
  - You and your biases

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

Consider now the questions you stipulated in the RUPAS evaluation from Sections O and P that might have involved the acquisition of qualitative data. Let's assume that the evaluation time line and budget allow you to collect the following:

- Interview data to better understand how RUPAS Program activities were designed and implemented along with the challenges and successes that program staff have experienced to date.
- Focus group data to shed light on how Parent Leaders and their families experience the RUPAS Program.
- Observation data via home visits so that you can watch how parents use the RUPAS curriculum with their children.

If you are the data collector and analyst, in what ways do the similarities or differences of your previous experiences influence how you view and understand the lived experiences of these different stakeholder groups as reported through interviews, focus groups, or field observations? In the event that another member of the evaluation team, a stakeholder, or someone who is more intimately involved with the program is charged with examining this data, how might her or his views affect the data analysis?

With respect to the development of codes and coding schemas, consider the ways in which personal assumptions about others' self-report experiences in particular support or hinder this process. How might they influence the identification of themes? What are some possible ways in which you can account for these influences during the analytic process? And what are the advantages and disadvantages of personally analyzing these data versus working with program stakeholders?

Finally, as we considered in Section R, how complex are the analyses that will be conducted? Are you able to arrive at results using a widely available tool such as Microsoft Excel as you would with a license-based package such as NVivo or subscription-based tools such as dedoose? How does the use of tools such as these affect the evaluation budget in particular and the evaluation in whole?

Resources

Atlas.ti
www.atlasti.com/de/index.html

dedoose
www.dedoose.com

NVivo
www.qsrinternational.com

Further Reading

Goodwin, D., & Rogers, P. J. (2015). Qualitative data analysis. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of practical program evaluation* (pp. 561–595). San Francisco: Jossey-Bass.

This chapter provides a comprehensive discussion of qualitative data analysis with several applications that demonstrate how the methods were used.

Henderson, S., & Segal, E. H. (2013). Visualizing qualitative data in evaluation research. *New Directions for Evaluation, 139,* 53–71.

This article offers some issues and guidelines to consider when analyzing and representing qualitative data.

Lyons, J. (2016, November 9). Qualitative chart chooser [Web log post]. Retrieved from *http://stephanieevergreen.com/qualitative-chart-chooser.*

This is a great tool for ideas on how to visualize qualitative data. Additional resources regarding data visualization principles are also available on the site.

Patton, M. Q. (2003). Qualitative evaluation checklist. *Evaluation Checklists Project.* Available at *irantvto.ir/uploads/qec.pdf.*

This is an excellent checklist of things to think about in qualitative evaluation.

Ryan, K. E., Gandha, T., Culbertson, M. J., & Carlson, C. (2014). Focus group evidence: Implications for design and analysis. *American Journal of Evaluation, 35*(3), 328–345.

This article offers a discussion on the implementation of focus groups in various programmatic settings. As analysis and collection of qualitative data tend to occur iteratively, this would be a helpful resource to consult when working within the confines of focus groups specifically.

Quick Reads

1. Eric Barela on Providing a Detailed Description of Qualitative Inquiry Choices and Processes to Clients
   http://tinyurl.com/hscownv

2. Michael Quinn Patton on Practical Qualitative Analysis
   http://tinyurl.com/hxfeb2c

3. Rahel Wasserfall on the Power of the Dissonant Story
   http://tinyurl.com/hmtvyr8

SECTION

T

How Are Analyzed Data Used to Answer Questions?

In Sections R and S, I talked about how to analyze quantitative and qualitative data, respectively. In those discussions, I focused on the analysis and presentation of data related to individual questions. Is presenting findings alone enough? Most evaluation writers would say "No!" Their view is that findings alone need to be supplemented with a *specific answer* of "good" or "bad." They would ask about the bottom line related to the data findings for each question and would want to know whether what you found was acceptable or not. Some evaluation writers, moreover, would further insist that judgments about individual questions alone do not suffice. The essential issue, they would maintain, relates to judging the merit and worth not only of each outcome but of the program as an entity. Their view is that the evaluator has the *personal* responsibility for providing a final judgment of merit and worth (e.g., the program is good or the program failed). *I do not agree.*

DIFFICULTIES IN VALUING

Aside from the issue of the appropriateness of this theoretical stance, let me ponder with you some of the difficulties involved in "valuing." How positive do the particular findings of a question have to be in

order to warrant a statement of "good"? Or, conversely, how negative to be "bad"? Are criteria sufficiently well established to be able to make such judgments? What if the answer to some of the questions is "good" and to others the answer is "bad"? How are judgments to be made about the total program given these diverse values? Are some questions more important than others? How much more important? Does the evaluator decide? How is this decision made?

These are daunting questions indeed. At this point I raise the further issue of the conditions under which making value judgments about the program are, in fact, *necessary*. Moreover, I ask you to consider differences in the nature of value judgments between summative and formative evaluations. Summative evaluations are generally designed to answer questions that lead to "go/no-go" decisions—that is, they provide the answer to whether the program is good or bad—what is its merit and worth? Typical questions include: Was the program successful? Is this program worth keeping? Should the program be refunded? Should the program be expanded to other locations? Judgments of merit and worth by the evaluator would certainly seem to be appropriate for summative evaluations, and further, it would seem to be appropriate that these judgments come from an external, unbiased source—the evaluator.

Formative evaluations, on the other hand, are designed to provide information for potential program improvement. It is anticipated that the program will continue and the evaluation is intended to provide insights into what is working and what is not, in order for modifications to be made. Formative questions might also seek to answer whether specific program activities had been implemented and, if not, what program modifications or changes might be appropriate. More broadly, summary formative evaluations would seek to take stock of the program's current status at some defined point in time (e.g., end of year); such evaluations have an orientation and intent of determining areas in need of further improvement.

I believe that most of what we engage in as evaluators are formative evaluations. They may focus on an ongoing process or on implementation questions, or they may deal with interim outcomes. In some instances, summary formative evaluations are conducted. In the total scheme of things, there are very few *real* summative evaluations conducted—and certainly not very many when we consider evaluation in small- to midsize local programs. My view is that in many cases, formative evaluations of such programs do not require that a valuing judgment be made, and if so, not necessarily by the evaluator. This is different from what I view as appropriate for summative valuing.

At this point, you might ask:

> "Marv, I seem to remember that in Section A, you talked about how you and your wife approached the purchase of a house. You came up with criteria that would be important. You had weightings that you attached indicating the relative importance of each criterion. Judgments were made on the criteria about each of the house candidates. Thus, while you asked separate questions (e.g., number of bedrooms, quality of bathroom), you managed to come up with a *single judgment of merit and worth* for the objects being evaluated. How is this different from the kinds of evaluation situations in which I might be involved?"

Good question. But you see, that was really a summative question. My wife and I were trying to decide, "Shall we buy or not?" In fact, it was a summative evaluation (nonprofessional) because we were comparing different alternatives about which we would make a summative decision. We were going to decide *which* house to buy. However, we, as evaluators, did not make the judgments—in that situation, we were the primary stakeholders. We were functioning as evaluators and helped ourselves as primary stakeholders to set up a valuing system to assist us in making a summative decision.

In contrast, formative evaluations related to the house situation would have focused on gaining understanding of the status of various aspects of the house. What things are working and which are not? Is the paint looking shabby? These are formative appraisals unlike a decision to buy or not to buy the house.

VALUING IN A FORMATIVE CONTEXT

So what kind of valuing, if any, takes place in formative evaluation situations? Let's take this one step at a time. First, given the nature of formative evaluation, a judgment about the goodness of the program as an entity is not required. It is often superfluous. The question of aggregating value judgments related to individual questions is not necessary.

What then is appropriate? Some evaluation writers maintain that even attempting to come to a final judgment about individual questions is inappropriate. They say that the values to be found within the data related to a question are subject to the individual perspectives of multiple stakeholders. They would maintain that we all see value from our own frame of reference. They would prefer that data related to

questions be provided in expansive description (they would say "thick description")—in that way, stakeholders should be able to make up their own minds based on this thick description. Thus, valuing from that theoretical perspective is not the job of the evaluator, but honors the subjective judgments of multiple individuals. There are further variations within that perspective, but I do not discuss here the fine-tuning that appears in the literature.

I don't go that far—that is not the tone of this book. Rather, I view the evaluator's role as much more *facilitative and enabling* in the valuing process. I seek to encourage stakeholder participation—I believe that such participation enhances "buy-in" by stakeholders—and, hence, increases the possibility of use. This is especially true with respect to valuing. If stakeholders set the standards for merit and worth, then it is more difficult to reject results and not to make changes when the findings are shown. In short, I seek to establish a *user framework for judging results*. Recall that in Sections J and N, I suggested that you work with primary stakeholders in getting them to depict more fully the parameters of acceptability related to each question—that is, what were their views (before any data were collected) as to what would be an acceptable set of findings for each question? Valuing was in large part performed by primary stakeholders before they might be biased by examining actual findings. As you might recall, I asked you, as the evaluator, to introduce scenarios of possible findings to determine acceptability from the perspective of these stakeholders. Where qualitative instrumentation was likely to be used and in other situations where quantitative prevaluing was not applicable, I urged you to inquire as to what they thought a successful program would look like. A qualitative or descriptive statement could then be used as a basis for subsequent comparisons. I suggested that you try to help them to write up a description of what success would look like. Am I fantasizing? In part, yes. This is an ideal-world description, which is less frequently met than I would hope for. It is difficult to get stakeholders to sit down and engage in this process, but I try, and sometimes I find success.

So in many instances, I settle for somewhat incomplete standards for judging quantitative findings and incomplete descriptive statements for questions using qualitative responses. While I view a complete description as important in potentially enhancing the ability of stakeholders to improve their program, that is not enough. It is my belief that description does not obviate the role of the evaluator in summarizing data and indicating patterns that are clearly found. Summary findings are not value judgments; they simply enhance the possibility of determining whether values previously specified had been

met. Where *means of judging value* were *not previously specified,* summary findings enhance the ability of stakeholders to determine the value of their program for themselves. Moreover, I like to guide the primary stakeholders' valuing by offering alternative conclusions or paths of action that might be warranted by the findings. These are presented as questions for stakeholders' consideration. The valuing activity, as I perceive it, is part of a process that encourages stakeholders to be users.

However, there are some situations where it is possible and appropriate for the evaluator to make judgments (good/bad statements). In those instances, questions might have been asked so specifically that the evaluator is able to answer the question in a yes/no (attained/not attained) fashion. An example of this is when the question itself had indicated a specific standard for judging (e.g., success as indicated by a score of 23 on the XYZ test). Alternatively, standards might have been preestablished for some questions answerable by using a standardized test. In that instance, the determination of "good" might have specified a particular norm level to be achieved. Even in these instances, there is a role for the evaluator in discussing the adequacy of that norm for future action.

There are perhaps more "sophisticated" ways in which valuing might take place. Typically, they are not warranted in formative evaluations. For example, evaluation in a summative context most frequently would involve the demonstration of value by showing the statistical significance of differences between outcomes of a program and its comparison program. (You might want to reexamine the discussion in Section P about causal models to note the complexity of randomizing or of selecting an appropriate control group. Section M is also helpful for this.) If, however, you had conducted the study using control groups, then significant differences between the two programs might provide a satisfactory indication of *merit*. Note, however, that *worth*—the second aspect of value—is not necessarily demonstrated if the costs of the two programs were different or if there were other aspects of the context not considered. In essence, the program differences, while *statistically significant,* might not be *practically significant* because of other factors unique to that context.

In Section X, in discussing utility models, I provide a detailed example of how cost considerations can be examined in programs with single or multiple outcomes. You may think of this as a much more detailed and professional version of the weighting system that I used in my house purchase example provided in Section A. You should note that even if the cost issues are not included, the procedure provides a more detailed and complex discussion of what I have proposed here.

➢ ***The Bottom Line:*** Valuing based on data from randomized control trials may be of great value for summative evaluations. This is particularly true when these "significance" data are combined with a weighting system for considering multiple outcomes. However, for formative evaluations within a framework focused on evaluation use, standard setting by primary stakeholders is the appropriate method for valuing. As previously noted, this stakeholder standard setting is informed by evaluator representations to primary stakeholders of the value positions of underrepresented stakeholder groups.

"VALUING" REDEFINED

It is clear to me that the initial description of valuing presented in Section A is flawed. That description implied that determining merit and worth was the responsibility of the evaluator. That is one view. Yes, evaluators can perceive their role as *personally* making a decision of merit or worth (see Alkin, Vo, & Christie in the Further Reading for this section). I maintain that there is no singular way of portraying the role of the evaluator in valuing. There are a variety of evaluator approaches and each carries with it different implications for the way that valuing transpires. Evaluators can be engaged in valuing by guiding stakeholders in the process of reaching conclusions about value. Evaluators can be engaged in valuing by acting as a social conscience in reflecting on the meaning of findings. Evaluators can assist in valuing by providing stakeholders with the opportunity to actively engage in evaluation, and in that process, themselves determine the worth of an enterprise. Based on what we have now seen, let me redefine the role of the evaluator in valuing. It is my view that *the role of the evaluator in valuing is to activate, be engaged in, and contribute to the process of determining merit or worth.*

➢ ***An Off-Beat Comment:*** Have I ever done valuing (of the determining good/bad variety) in a formative evaluation? Well, yes. Let me explain. Many years ago, I was engaged in the conduct of an evaluation of the educational programs at all of New Mexico's juvenile detention camps. My stated value judgment at one of those facilities was "This is an educational disaster area." I would say that was a value statement that went well beyond simply saying "bad."

Now, how did I get there? This small facility, housing about 25 youth, was located on the highest mountain in northern New Mexico. It had one classroom and was allocated one teacher. When I arrived at the facility

there had been no teacher obtained and the classroom was not in operation. The staff had done nothing to compensate for these deficiencies. The juveniles mostly hung around the large, lodge-like building all day. When staff were questioned about any kind of educational program or learning activities that might be available, none could be mentioned. One staff member indicated, hopefully, that they sometimes let the youth go out to chop wood. Obviously, the state's newspapers had a field day with my "disaster area" quote. Fortunately, before the criticism got out of hand, an influential member of the state legislature who had come along as an observer seconded the value judgment made. So when it comes to whether I personally assign values in an evaluation, I can never say "never." In this situation, the evidence was so incontrovertible that I, as the evaluator, was thrust into making a value statement. Might it happen again? Possibly.

RECAP—SECTION T

Valuing

- Traditional View of Valuing
- Difficulties in Valuing
  - Formative–summative differences
- Valuing in a Formative Context
  - Developing a framework for judging results
  - A comment on summative valuing solutions
  - The bottom line
- Valuing Redefined
- A Final Note

GAINING ADDITIONAL UNDERSTANDING

Thought Questions

In most cases, data are collected and analyzed with the intention of using that information to arrive at new insights and inform decision making. With the evaluation that you have been considering in mind, think about whether you have adequately engaged primary stakeholders in prespecifying criteria for determining "value." If you did this, then how might you work with stakeholders in applying those criteria? Be sure to reinforce the idea that this is their valuing system. If you did not do an adequate job of establishing a valuing framework, then what could you do now to involve them in valuing in an unbiased manner?

Further Reading

Alkin, M. C., Vo, A. T., & Christie, C. A. (2012). The evaluator's role in valuing: Who and with whom. *New Directions for Evaluation, 133,* 29–41.

In this article, my colleagues and I discuss the various evaluator perspectives on valuing.

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. *New Directions for Evaluation, 68,* 15–32.

You will find, in this article, a thoughtful analysis of the ways in which one could arrive at various evaluative claims that are grounded in evidence and critical reasoning.

Julnes, G. (2012). Managing valuation. *New Directions for Evaluation, 133,* 3–15.

This article is an extension of Fournier's analysis, but with a more intentional focus on current valuation practices. It is one of the few resources available on valuation and not to be missed.

Patton, M. Q. (2012). Contextual pragmatics of valuing. *New Directions for Evaluation, 133,* 97–108.

As you know by now, context is key in evaluation. Patton underscores this point in his paper as he illustrates the ways in which context-sensitive valuing occurs across different settings.

Quick Read

1. Bonnie Richards on Valuing in the Social Services
   http://tinyurl.com/hp4dj3m

SECTION

U

How Are Evaluation Results Reported?

Evaluation has, as an ultimate purpose, the improvement of the program being evaluated. This is especially true for formative evaluation. When evaluation results are reported it helps to ensure that relevant information is available so that it may become part of the process of considering ways to potentially improve the program. Thus, there is a necessity for paying heed to the way that you report to stakeholders. However, let me first point out the relationship between reporting and communicating.

COMMUNICATION

Evaluation reporting typically means the reporting of findings. These findings can either be presented as the program proceeds and the evaluator acquires data, or at intermittent points throughout the evaluation—particularly at the end of the program year. But reporting is part of a larger entity, the act of communicating, which occurs throughout the evaluation. Indeed, the very act of conducting an evaluation involves communication at every stage—two-way communication—this, surely, must be obvious at this point in your reading of *Evaluation Essentials*.

Evaluators identify stakeholders and engage them in order to determine their information needs. Stakeholders communicate their views and their concerns; evaluators communicate as well. Often, the communication and information provided by evaluators helps stake-

holders to properly identify their most salient information needs. The development of an evaluation plan involves a good deal of communication as evaluators inquire about stakeholders' needs and indicate what is possible. A simple reading of all of the proceeding sections provides abundant evidence of an active communication process that encompasses all stages of the evaluation. Continuous communication of all types is an essential part of the evaluation process; your communication throughout the evaluation process may provide insights and understandings that will lead to program improvement. This we have referred to as "process use"—the impact of the process of conducting the evaluation on decisions, other actions, and understanding.

ISSUES RELATED TO REPORTING

"Evaluation reporting" refers to the specific set of information provision activities in which evaluators communicate what it is that they have learned about the program, based upon their systematic study. This reporting can be thought of in several segments: reporting that occurs *throughout* the process of the evaluation and reporting that occurs *at the end* of the program year or at other designated points in time.

Ongoing Reporting

Let us deal with the first of these—ongoing reporting—and look at several relevant issues. The main issue here relates to the insights you gain about the program during the course of the evaluation. Such insights might, for example, relate to the extent to which the program is being operated in the fashion in which it was intended. You, as the evaluator, might note, for example, that some of the activities have not taken place or have taken place in a poor manner. When conducting interviews or observations you might begin to develop insights into increased negative attitudes toward the program. I believe that it is inappropriate in most instances to allow the program to proceed on a course toward inevitable failure. I feel that you, as the evaluator, have an obligation to *report* such *interim findings* as they occur. Let me caution, however, that there is a fine line to be drawn in making these decisions about when to report. There are hazards involved in providing extensive information prematurely that may not be accurate. In such instances, you should state the tentative nature of this reporting.

There are many ways in which interim evaluation findings are reported. You might prepare written "mini reports," describing what has been learned to date, or there can be more casual reporting, includ-

ing such things as e-mails, oral reports, telephone reports, or visual presentations.

Let me restate the broader issue involved: I view continued engagement with stakeholders (particularly primary stakeholders) as a key part of the evaluation process. Thus, ongoing reporting of your observations and impressions is essential. This is especially true for those evaluators who have a focus in their evaluation orientation toward organizational learning and evaluation capacity building.

Reporting Final Results

There are also a number of issues to be addressed related to final reporting. Most end-of-evaluation findings involve a written report—however, that is not always the case. The first issue to be addressed is: *Are final reports necessary*? Some evaluation authors have questioned the cost-effectiveness of written reports—that is, would the program stakeholders want to use the funds that would be necessary to produce a written report for other purposes, such as enhanced program activities or additional evaluation work? Report writing is a very labor-intensive activity. Would a less costly presentation to all relevant stakeholders be equally effective at a substantially lower cost? Perhaps such a presentation could be videotaped so that many others could have access to the report. Does the ongoing reporting that is to occur suffice? Nonetheless, written reports are the most common form of final reporting. (Note that when I talk about final reports, much of what we have to say also refers to other scheduled reports—such as a midyear report.)

I would like you to recall that in Section D, we discussed some necessary agreements that must be reached about things such as timing of the report and format. Final results can take on a variety of formats. *Stakeholder format and information detail preferences* are an important consideration. Some people find certain kinds of reports or formats more comfortable and more easily understood. There are a variety of reporting formats that seasoned evaluators might employ—multimedia, electronic reports, websites, storytelling, sociodramas—to name but a few. And, of course, there is a written report. Then, there is the question of the extent of detail that is appropriate in a report. Different stakeholders have different levels of information needs and report-reading tolerance. Thus, reports might need to be presented at multiple levels of detail. This can often be accomplished by proper formatting of the evaluation report. I elaborate on this later in this section.

Another issue that is usually not addressed but is very important relates to how and by whom the report may be cited. I note that some evaluation writers talk about the need to establish "ownership" of the

data and the report. One concern that they state is the right of the evaluator to unilaterally decide to publish the evaluation findings (e.g., in a journal). I do not feel great concern about this matter because of my very strong commitment to evaluation use and, therefore, to the unique relevance of the evaluation for the particular program site being evaluated. The evaluation was conducted to learn about *this* program in *this* context at *this* time. If there is something in the evaluation that is sufficiently generalizable and of interest to a broader audience, it is the client's right to determine whether that will be allowed. However, the right of client ownership has its limits. It does not extend to, in any way, modifying the evaluator's report. In my view, instances of doing so, or of the use of inappropriate or misleading quotes from the report, authorize evaluators to make their public voice heard. I talk about this somewhat in Section V where I address the issue of evaluation misuse.

A main issue in reporting is to ensure that the evaluation report considers the *needs of multiple stakeholder groups*. Do some groups require reports presented in a language other than English? Are there diverse communities that are not normally part of the information network for which particular reporting should be devised? More broadly, the issue involves reporting in a culturally sensitive manner. The idea of an evaluator's cultural competence has been noted in earlier sections of this book. There are many ways in which it is reflected in evaluation reporting. Perhaps my general advice to you is to recognize the way in which your communication and language understand, reflect, and respect cultural differences. Accomplishment of this goal may require multiple reports and reporting methods. How are *reports distributed*? I have discussed previously in this section how the way in which the report is formatted might serve the needs of multiple audiences. Furthermore, in disseminating the evaluation report, you, as the evaluator, might plan for it to be distributed in pieces, providing only those sections appropriate to each audience. Thus, there might be many copies of executive summaries provided to various stakeholder groups and only a limited number of the full report, including the technical appendixes. There might also be the possibility that reports are presented in other languages. These are issues to be resolved with the primary stakeholders.

THE FINAL WRITTEN REPORT

Now let us turn to the discussion of a final written report. I will summarize my discussion in two categories: the format and structure of the report, and the nature and quality of the writing itself.

Elements of the Report

With respect to the first of these, I view a typical report as consisting of an *executive summary,* a *brief description* of the program, the discussion of the evaluation *procedures or methodology,* the *findings,* and finally, an *appendix* of notes or attachments. Some evaluation writers add evaluator recommendations to this section—or include the recommendations within the findings section. I later discuss my views of the appropriateness of such an inclusion (in either form).

Let the end be the beginning. Confused? I strongly believe that the *executive summary* should be the very first part of the report. We, as evaluators, must, at the very outset of the report, succinctly address the stakeholders' concerns: What did you find? The executive summary should provide a brief description of the program that was conducted. Next in line, a brief description of the methods and procedures used by the evaluator to acquire understandings about the program. Then, and most importantly, the questions and the main findings associated with each question needs to be briefly summarized. Finally, the appendix provides further detail.

I believe that the executive summary provides an introduction to the report that allows readers to first gain the full picture very quickly, which is often the only thing busy administrators will read anyhow. The individual sentences or brief paragraphs in the executive summary should provide an appropriate overview of the findings. Each major finding should be accompanied with a page number indicating the location in the report where that particular finding will be addressed more fully (see Figure U.1). Thus, some readers may only read the executive summary and from that gain a general understanding of the findings—the bottom line. Other readers of the executive summary will note the page references and be motivated or inspired to read, in greater detail, about the findings that are of special interest to them. More determined readers will desire to read the whole report including program description and methods. Technically astute readers will also avidly consume the appendix to the report.

Following the executive summary I would place a *description of the program* itself. Simply stated: What was the program that was implemented? What were the goals of the program? What major activities were pursued in an attempt to accomplish those goals? If the program details are quite complex, then I suggest that they be summarized and readers then be referred to an appendix.

Next is the discussion of the *evaluation procedures and methods.* These would have been fully explicated in the evaluation plan that you developed with stakeholders; that plan can be summarized here. How-

Program

_____

_____ See pp. _____ to _____

Methods

_____

_____ See pp. _____ to _____

Results

Question 1: _____

Findings: _____

_____ See pp. _____ to _____

Question 2: _____

Findings: _____

_____ See pp. _____ to _____

Question ____: _____

Findings: _____

_____ See pp. _____ to _____

Unintended consequence/overlooked area: _____

Findings: _____

_____ See pp. _____ to _____

FIGURE U.1. Executive summary.

ever, there may have been procedural or methodological changes that arose during the conduct of the evaluation—these changes should be noted. In describing procedures, it is important to remember that part of the procedure includes the process that you, as the evaluator, went through, including such things as developing a logic model and framing the decision questions and scenario testing for relevance. While research reports tend to require substantial discussions of methodology within the body of the report, evaluation reports are different. I feel that in a decision-oriented situation such as an evaluation, where users have been involved from the outset, it is not necessary to encumber readers with overly extensive detail of procedures as part of the main body of the report. This section should provide adequate detail so that the reader can understand the evaluator's approach. However, all of the complexity does not need to be described. A further description

of particular aspects of the methods, along with copies of the instruments used, should be placed in the appendix.

And now to the *findings*. The findings constitute the bulk of the report. Findings, of course, should be keyed to the particular questions or issues that were determined in the plan as being important to relevant stakeholders. That means that the questions should be the foci of each portion of the findings section. Data for answering a question may have come from a variety of data sources—both quantitative and qualitative. The evaluator's job is to integrate these data into sensible answers to questions. Data should be presented by questions. A brief summary of data derived from each particular instrument and the analysis of these data could initiate the findings section. Then, however, the focus must turn to answering stakeholders' questions. What is the case being made by the data? What does the full data set mean? Do different subsets of data (qualitative and quantitative) provide corroborating or contradicting findings? Often, qualitative data are an effective means of elaborating and exemplifying quantitative findings.

Frequently, issues emerge that were not a part of the initial question set—that is, they were *overlooked*. These might be questions or issues that later were felt to be of interest by the primary stakeholders. In that instance, they should be considered along with the initial questions. Sometimes, however, the evaluator is confronted with data indicating *unintended consequences*—either positive or negative. These should be duly noted and addressed within the findings section. As noted in Figure U.1, these are included in the executive summary, as well.

Needless to say, despite our best intentions, not all questions are able to be answered, or, at least, not all questions are able to be answered definitively. Evaluators cannot, and should not, go beyond the data that they have. Findings need to be well substantiated and where they cannot be—say so. You can only report what you know—what you have learned. Do not go beyond that.

Furthermore, not all findings are positive. Clearly, you, as the evaluator, must resist the temptation to provide pleasure (good news) when it is not warranted. Bluntness, on the other hand, is not necessary. There are judicious ways to communicate *unfavorable findings*. Information to stakeholders implies suggestions for improvement. Thus, framing negative findings as opportunities for potential improvement is fruitful behavior. Also, to the extent to which you have established nonthreatening relationships with stakeholders throughout the evaluation, negative findings are more easily accepted. You, hopefully, would have become a trusted advisor. Needless to say, none of the above should be done in a way that damages your integrity. Presenting findings in

a sensitive manner is important, but an evaluator's integrity must be preserved.

Do I advocate making *recommendations* part of the findings or not? Many evaluators view recommendations as an inherent part of an evaluation. I take issue with that point of view. (However, you may feel free to disagree and go the way of the majority.) Now let me justify my point of view. I believe that findings point out either strengths or deficiencies. If the findings are positive, the implication for program continuance in its current form might, thus, be implied. (However, that is not necessarily so.) Furthermore, if the findings are positive but also show particular tendencies that might lead to potential modifications or enhancements of the program, then there are choices to be made. What is the best possible way to do this? For example, let us say that Activity A seemed to have a particularly positive impact. Should the program do more of it? How much? At what cost? These are decisions that I believe are administrative in nature. The evaluator is an expert in systematically collecting, acquiring, synthesizing, and helping to value data. Evaluators are not necessarily expert in the content of the program, but many, if not most, of the relevant stakeholders are. Evaluators usually know less about the content and goals of drug prevention programs, welfare-to-work programs, or mathematics programs than the stakeholders. Can the evaluator assist in discussing possible alternative recommendations? Can the evaluator assist in explaining what the data have to show or imply about each of these potential courses of action? Sure, and sure. I frequently pose several possible courses of action that might be considered. These are presented in order to stimulate thinking about what might be feasible and appropriate.

I, however, like to *assist* in the recommendation process *up front*. Up front? Yes, *way up front*. When the evaluation questions of importance are being determined and I ask whether stakeholders really want answers (remember those earlier sections?), I like to put forth the issue of "What difference would it make? What would you do given this, that, or another set of findings?" In essence, I am asking the primary stakeholders to consider what recommended courses of action might be implied given particular future findings. I am asking for a consideration of potential future recommendations given possible outcome results. The examination of potential recommendations is part of the evaluation focusing and planning (particularly in Sections J and N).

Now, finally, let us turn to the last portion of the written report and consider the *appendix*. I have indicated at previous points in this discussion the various materials that might be placed in an appendix or attachment to the report. The intention is to keep the report within a reasonable page limit and to enhance its readability. The compulsive

stakeholder, or perhaps the more sophisticated stakeholder, will want to read everything. And, as noted, there are intermittent stages for those who want only a summary of the findings (the executive summary), or the findings with additional detail about how those findings were reached. The appendix is the place to provide all of the substantiation of the report. If statistical methods were used, detailed discussion of the analysis procedures and the more esoteric analyses find their home here. Copies of the instruments rest here as well.

RECAP 1—SECTION U

Communication and Reporting

- Communication
- Reporting
  - Issues related to reporting
    - Ongoing reporting
  - Final reporting
    - Final report: necessary?
    - Final report: ownership
    - Final report: stakeholder format preferences
    - Final report: multiple stakeholder groups
  - Final written report
    - Elements of the report
      - Executive summary
      - Program description
      - Procedures and methodology
      - Findings
      - Recommendations?
      - Appendix

Quality of Presentation

While content is important, it must be communicated in a way that is understood and persuades. The first thing to be said about the quality of writing is that evaluators must write for the audience. Who are the potential readers? What would be readable? In my writing, I like to think about a moderately informed, but not fully knowledgeable stakeholder. I then picture this reader and, as I write the report, ask whether

that individual would understand what I am saying. When I say something that I think might be difficult to comprehend, I explain it again in another way or give an example. Think of a person you know and imagine that you are talking to that individual, but also imagine communicating in a culturally sensitive manner.

For a report to be readable, it must have a *clear and accessible style*. What do I mean by "accessible"? You should write without an abundance of complex syntax. The writing should not be convoluted. Simple sentences without a lot of "add-on" thoughts are preferred. I may be guilty of violating this rule in the current book (since I seem to be very fond of semicolons and dashes). The evaluator might also want to vary the length of sentences to avoid monotony. Sentences all of the same length, style, and tone are not interesting to read.

Let us, jointly, take a moment to reflect further on one of the most serious problems with professional writing—*jargon*. By jargon I mean primarily the technical language of evaluators—the terms that I have tried to avoid throughout this book (and which are far more profuse in most other evaluation texts). Evaluators too often feel the need to explain things in technical terms when they could be explained more simply. (This really is like a form of bragging to demonstrate one's sophistication.) I suppose this applies to everyone and to all situations. Some technical terms may be necessary. If you must use them, explain each simply upon first use.

One can write and indicate technical details and still do so in writing that has a pleasing *verbal quality*. A writing style that is mechanical or plodding does not add to readability—the aim of a report is to be convincing. People cannot be convinced by something that is boring, and they end up not having the patience to read to its end. Think of the novels that you have started reading and put down uncompleted because they were too boring. You were not convinced that the author had anything of interest to say. You want the readers of your evaluation report to be convinced that you have something worthwhile to say.

Furthermore, it is important to *guide the reader* in understanding what you are about to say. You want the reader to understand the flow and sequence of the argument to be presented. Executive summaries, as we have discussed them earlier in this section, provide an overview of the full report and help the readers get into it and to understand its structure. Your next job is to guide the reader in dealing with the content in the body of the report. Guiding paragraphs that set the framework for a section are helpful. They introduce the reader to what is to come. You may also guide the reader in reading individual paragraphs. Many writing specialists maintain that the initial sentence of a paragraph is important because it sets the tone and structure for what is to

be further detailed within that paragraph. Often, a well-written report can be roughly summarized by simply reading the first sentence of each paragraph.

Carefully chosen *illustrations* help, for example, to demonstrate the point being made within a paragraph or portion of a report. Perhaps it is simply noting a striking statistic or repeating memorable quotations from the qualitative data. Possibly, you might use an example from another, more easily understood, domain of the concept being addressed. Sometimes metaphors help people to understand the message being conveyed. Metaphors are persuasive. The one caveat to be noted is that illustrative examples generally should be representative of the general trend of the data and not of the outliers.

I have provided some guidelines to good writing, some of which have been detailed, but realize that *writing is also an art*. Practice it, criticize yourself, and let others comment. Listen to comments, but your writing should reflect you and what you are comfortable saying and the way in which you are comfortable doing so. That is a key to successful evaluation reporting.

DATA VISUALIZATION HELPS

There are now available an incredible array of data visualization techniques. These techniques, if used correctly, can aid reader understanding. The simplest and most long-standing of these are such things as *tables, charts, maps,* and *other graphics*. The computer program Excel is a handy way of converting quantitative findings into tables and graphs. Of course, these must be used judiciously in ways that add to, rather than detract from, the report. Overly abundant, overly complex tables are not helpful—the key is to prioritize the most relevant information to help improve clarity and understanding. For example, simplified data tables may be more useful in the main body of the report, and detailed tables can be placed in the appendix to supplement the readers' information needs.

In recent years, the use of graphic devices in evaluation reporting has grown in popularity and is commonly referred to as data visualization. Whereas before it was sufficient to write a detailed report, now stakeholders expect to see striking visualizations that aid interpretation of findings. High-quality visuals can help spotlight key messages, patterns, and observations to inform stakeholder decision making (aka evaluation use).

The most commonly used tool for tracking quantitative data is Excel, which is included in the Microsoft Office suite. It provides a good

way of quickly exploring data or creating visualizations and is fairly easy to use with good documentation. Users can modify the normal (default) setting to create professional graphics. Another useful tool is called Tableau. This is available in either a free Web-based public version or paid desktop subscription.

A good graphic is easy to understand and interpret. It takes some practice, but everyone can make visually compelling and easy-to-understand graphs and charts. First, briefly summarize the key finding in the title. If additional context is necessary, add a subtitle or a brief paragraph near the graphic to aid comprehension. Next, simplify your graphic. Remove unnecessary grid lines, backgrounds, tick marks, and text. Use color and arrows to emphasize only the data that are connected to the key message. Figure U.2 is a graph produced through Excel that was modified by the removal of grid lines and tick marks, and a title was written to draw stakeholders' attention to the key finding.

While data visualization is most commonly used with quantitative data (e.g., graphs), it can also be helpful when sharing (or even analyzing) qualitative data. One way of portraying qualitative data is through a "word cloud." In essence, this uses a computer program to depict the frequency of word use. The more common a word is, the bigger and bolder that word will appear. Figure U.3 is a word cloud that represents the "observations" portion of Section L.

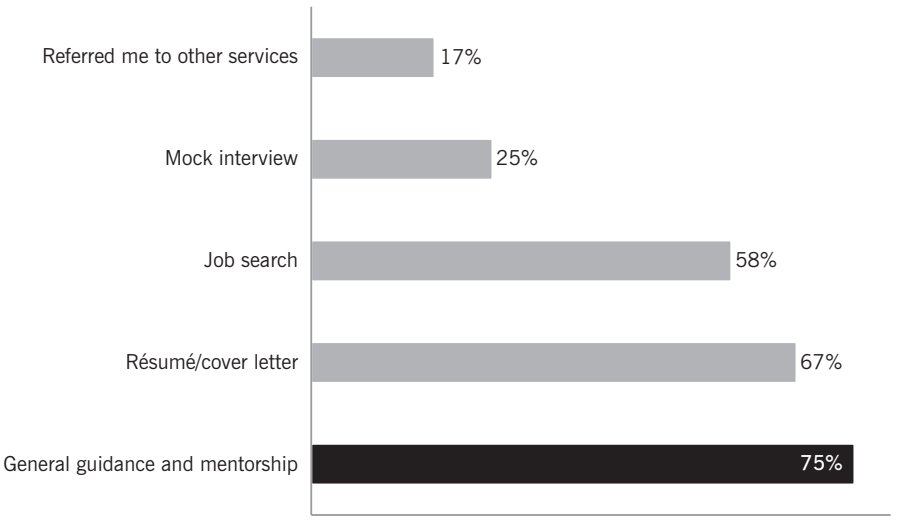

FIGURE U.2. Seventy-five percent of students rely on career counselors for general guidance and mentorship.

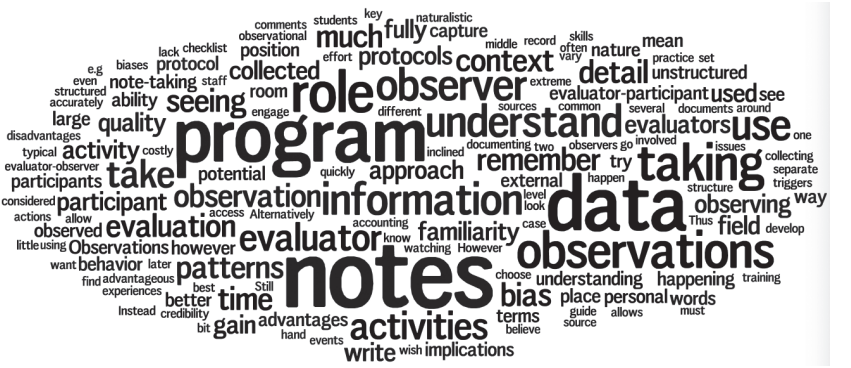

FIGURE U.3. A word cloud of the "observations" portion of Section L.

Data visualization is a growing field with new tools constantly being developed. While what I have discussed here is a start, be aware of new tools that might be easily employed.

Finally, a simple suggestion—value clarity above all other elements of style and content type.

RECAP 2—SECTION U

Quality of Presentation

- Writing
  - Focus on potential stakeholders/readers
  - Write in a clear and accessible style
  - Avoid jargon
  - Use pleasing verbal quality
  - Guide the reader
  - Use illustrations where appropriate
  - Develop your writing skills—writing is an art
- Data Visualization Helps
  - Use tables, charts, and so on where appropriate
  - Use more advanced data visualization techniques

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

As with other topics that we have discussed in this book, we come to issues of reporting and communication with an eye toward use (more on this in Section V). Thus, when thinking about how best to communicate your evaluation findings and developing a reporting plan, we encourage you to consider several issues. Who is the audience for your evaluation reporting? In the RUPAS context, is it program beneficiaries, Amy Wilson and her team, Family Matters, the Children's Trust, or some combination of these groups? How will they use the information that you share with them? Would you anticipate that a single evaluation report will suffice, or will multiple versions be necessary to accommodate different stakeholder groups?

Now, think about your findings and consider the most appropriate venue for relaying them. Have you shared interim findings through e-mail, phone calls, or in-person meetings? Are there some stakeholder groups that might receive an executive summary alone? Who should receive the full report? Likewise, what are the advantages and disadvantages of each of these modes of communication for your project? What kind of communication style or approach does each require?

Resources

Microsoft Excel
office.microsoft.com/excel

Tableau
https://public.tableau.com/s/#tour

WordCloud
www.wordclouds.com

Further Reading

Alkin, M. C., Christie, C. A., & Rose, M. (2006). Communicating evaluation. In I. Shaw, J. Greene, & M. Mark (Eds.), *SAGE handbook of evaluation* (pp. 385–403). Thousand Oaks, CA: SAGE.

This chapter is a rather complete discussion of the evaluator's role in communicating to stakeholders and includes informal communication, interim reports, and formal reports.

Evergreen, S., & Metzner, C. (2013). Design principles for data visualization in evaluation. *New Directions for Evaluation, 140,* 5–20.

This article discusses data visualization with the goal of communication, which consists of encouraging the audience's engagement with the data so as to increase understanding.

Johnson, J., Hall, J., Greene, J. C., & Ahn, J. (2013). Exploring alternative approaches for presenting evaluation results. *American Journal of Evaluation, 34*(4), 486–503.

Recognizing that reports, tables, and graphs are a limited set of possible means to report evaluation findings, this group of authors offer visual displays, performances, multiple program theories, and poetry as alternatively viable approaches for communicating evaluative conclusions.

Lysy, C. (2013). Developments in quantitative data display and their implications for evaluation. *New Directions for Evaluation, 139,* 33–51.

This article offers a discussion of several different ways in which quantitative data can be visualized, seeking to go beyond the traditional contingency table, bar graph, and line chart.

Quick Reads

1. Lauren Baba and Carol Cahill on Evaluation Writing for Community Partners and Other Audiences, Part 1
   http://tinyurl.com/hclcnom

2. Amy Germuth on Using Visual Design Principles
   http://tinyurl.com/jola3yp

3. Courtney Howell and Shelly Engelman on Social Psychological Tips for Effective Reporting
   http://tinyurl.com/zw8hphz

4. Jeanne Hubelbank on Using Metaphors in Evaluation
   http://tinyurl.com/zoacen2

SECTION

V

What Is the Evaluator's Role in Helping Evaluations to Be Used?

In this section, I provide some explanation about what I mean by "evaluation use" and the many forms that it takes. Furthermore, I discuss what you, as the evaluator, can do to enhance evaluation use. I consider this section extremely important because of my view that evaluation is not simply some kind of intellectual exercise resulting in a report. All too frequently, such reports are received, smiles and thanks are given, and the report sits on a shelf. I don't want that to happen. Reports, like books, are meant to be read. Good books inspire. I hope that your evaluation reports will not only inspire but will lead to changed thinking and actions within the programs that you evaluate.

Greater methodological rigor alone does not lead to increased use. As I have mentioned throughout this book, methodological appropriateness is important, but it is not enough. Beyond that, there are many things that you, as the evaluator, can do to foster evaluation use. You must play an active role in increasing the likelihood that use does take place—not only at the end of the process but throughout.

In that sense, this is not really the fifth to last section of the book. In fact, preparation for utilization starts at the *beginning* of an evaluator's engagement. To repeat yet again, the first step in helping to enhance evaluation use—that the evaluation has impact—and that your work does not go for naught, is for *you* to actively *commit yourself to use* as a goal. You must think about and focus on evaluation use throughout the process.

A WORD ABOUT "USE"

The word "use" has certain intuitive meanings that you are most certainly aware of. Before I provide a precise definition, I would like to set two boundaries on this discussion. First, when I talk about use, it is in relationship to a *particular program*. The concern that you must have, as the evaluator, is whether the stakeholders you deal with are in a position to use the evaluation to improve their program. The second boundary that I want to establish relates to *intention*. I am more concerned about use that was intended—that is, instances where you believe that there is a possibility that engaging in the evaluation process or receiving particular evaluation information will benefit stakeholders. There is often use that occurs that is incidental, unintended, not anticipated, or far away in another program. These kinds of use, while worthwhile, are less under the control of the evaluator. They are more the fodder of evaluation researchers and theorists who like to talk about such things. I want to focus on actions within the grasp of the evaluator that might lead to use—and to potential program improvement.

WHAT IS USE?

Evaluation use refers to the way in which the evaluation process and the information obtained from an evaluation impacts the program that is being evaluated. By this I am referring to such things as the following: Did the evaluation generate a new understanding of certain aspects of the program? As a consequence (or partial consequence) of the evaluation, were changes made in the program? Did engaging in the evaluation process lead to new awareness and insights or to changes in the program? Did program staff acquire new skills and insights during the course of an evaluation that were attributable, or partially attributable, to the evaluation?

There is a great deal of research and other writings about the topic of evaluation use. Indeed, it may be the area of evaluation that has been the most thoroughly researched. When discussing evaluation use, evaluation writers tend to make the distinction between "instrumental use" and "conceptual use." *Instrumental use* refers to situations where evaluation information has been used to impact on direct action such as making particular decisions about a program—that is, evaluation information was a direct instrument for making change. "Direct impact" does not imply that the evaluation was necessarily the only source of information. Local knowledge and beliefs may (and usually

do) provide insights as well. *Conceptual use* describes situations where no direct decision has been made, but where particular conceptual understandings about aspects of the program have been modified based on the evaluation information.

Evaluators also refer to situations where evaluation is employed to justify a prior decision. In these instances, the purpose of the evaluation was not to answer a question, but rather to ratify an action already taken. I do not consider this as an instance of real use. This kind of "symbolic" use is not part of our discussion.

Instrumental and conceptual use may take place as a result of two different aspects of the evaluation. Most thoroughly examined within the literature on instrumental and conceptual use is the use of *evaluation findings*. Indeed, evaluation use may occur as a consequence of such findings—that is, does the evaluation report eventuate in instrumental changes or conceptual modifications? When you, as the evaluator, conclude an evaluation and provide a final report, does that report influence potential changes or improvements in the program? Note, as I have reminded you throughout this book, that findings from interim reports may also be used.

I have also emphasized the importance of engaging the primary stakeholders—and to some extent, all stakeholders—in the process of evaluation. We have talked about the role of stakeholders in the development of logic models. We have addressed the role of stakeholders in selecting questions to be examined and developing a system to "value" findings. On a somewhat less active level, stakeholders may have observed (or participated with) the evaluator in developing instruments, conducting interviews, collecting data, and so on. These kinds of engagement and observation of the evaluation process may also lead to evaluation use. Stakeholders have learned something from being part of the evaluation process. What they learn may influence what they now know about the program, changes they might make, or how they now think about the program. Engaging in the evaluation process may also increase their appreciation for evaluation—and future receptivity to it in other program evaluations. It might also increase their knowledge of evaluation. Evaluation writers refer to these kinds of use as *process use*—use as a consequence of engaging in the evaluation.

Thus, let me summarize. I ask you to first think of evaluation use as occurring either conceptually or instrumentally related to the findings of the evaluation. Evaluation use may also occur because of engagement in the process; this use may be either conceptual or instrumental.

WHAT CAN YOU DO?

There is an extensive literature on the factors associated with increased evaluation use. For simplicity, I consider some of these factors in three time frames. These factors are discussed more fully in Section Y, but for now, consider first some fixed characteristics of the program and evaluator situations that, when they are present, are more likely to lead to evaluation use. Think of these as the *preconditions.* The second category relates to the actions that you, as the evaluator, might take *during the conduct of the evaluation* that are likely to increase evaluation use. Finally, I believe that there are evaluator activities at the *end of the evaluation contract* (and even beyond) that are part of the evaluator's responsibility for enhancing use.

Preconditions

Let us consider first the preconditions that, if present, are more likely to lead to evaluation use. Clearly, some of these are not controllable. Others are potentially modifiable. The first, and perhaps most important, of these relates to *you* and who you are and what you believe. One of the most important aspects of evaluator characteristics is your *personal commitment to use*—the commitment to attempting to enhance use. You need to strongly believe that evaluation is important. You need to want to really make a difference. I have talked about this in earlier sections of the book and hope that you have already acquired that commitment to use. I sincerely hope that reading this section further convinces you of the importance of having this commitment. Use is more likely to take place if the evaluator involved is perceived as a credible source of information. If you are viewed as credible, your evaluation will be more respected. This *credibility* initially is based on prior reputation and initial stakeholder perceptions. Part of credibility, obviously, is the evaluator's technical capabilities and understanding of evaluation, but, take heed, credibility is not only based on perception at the outset of the evaluation. Evidence strongly suggests that credibility is, in large part, acquired. Changes in credibility may come about when evaluators demonstrate their abilities and skills in the conduct of the evaluation and provide evidence that they listen, care, and are nonpartisan and unbiased.

Another set of evaluation use preconditions has to do with primary stakeholders. Stakeholders, and in particular primary stakeholders, have views about evaluation generally that might impact their engagement in the evaluation and potential use—that is, stake-

holders' may have previously been part of an evaluation. Those prior evaluations may have been an unpleasant experience. Perhaps a prior evaluator did not seek stakeholder engagement and, while conducting a technically sound evaluation, missed the mark on what stakeholders wanted and needed. Perhaps the findings of the evaluation were quite negative and were presented in a manner that did not enhance receptivity. Your job is to attempt to alleviate this negative perception of evaluation and evaluators by stakeholders. You are providing a new kind of evaluation, one that is more stakeholder sensitive and one that is concerned about providing assistance rather than simply judging. Let the stakeholders know this.

Consider also the larger organizational structure that encompasses the program. With respect to the larger organization, there may be forces or views that evaluation is not important. There may also be severe bureaucratic constraints. These are difficult to overcome but be aware of them and try, as best possible, to work around them.

During the Evaluation

An issue associated with evaluation use is the extent to which *stakeholders* have an *interest in this evaluation*. Do they even want an evaluation conducted? Is it being forced upon them? Is this something that they perceive as a rite of passage associated with the acceptance of a program development contract? If so, on any of these, they are not likely to be highly interested in the evaluation or its potential use. These are, in part, preconditions of the situation. Your role in fostering interest in the evaluation is an important part of conducting the evaluation. You want to get stakeholders to understand that this evaluation is different. This evaluation seeks to be of help to them in improving something that they care about—their program. In part, this is accomplished by the process that we have described in this book, starting with the development of questions that are relevant to stakeholders and including other activities that increase stakeholder "buy-in."

In the past, I have had contracts to do evaluations of programs that were externally funded. In those instances, I encouraged stakeholders to consider what would be minimally necessary to satisfy the evaluation requirements of the contracting agency. Then, having satisfied that, I attempted to turn the evaluation into something that focused on primary stakeholders' real evaluation concerns. In short, I sought to increase their interest in evaluation and its use.

Evaluation works best and has the greatest likelihood of use occurring where it is elevated to the status of *valued norm* of the organization—that is, the stakeholders with authority as well as others in the organi-

zation make evaluation a part of their *modus operandus*. It is the way that things are done. In such a situation, people look to evaluation as "what we do" and what we consider valuable. Obviously, this is a very hard goal to reach, but keep it in mind as part of what you want to reinforce throughout the process of conducting the evaluation.

I talked earlier about the importance of evaluator credibility and how credibility accrues throughout the process. I mentioned technical expertise and, of course, communication. Another aspect associated with building credibility and, in turn, enhancing use is *timeliness.* It is essential for you, as the evaluator, to respond quickly to administrative requests and to provide information and reports in a timely fashion. Changes in program operation do not occur instantly. Give stakeholders the time to digest evaluation findings so that action can take place. The extent to which the evaluator's behavior and actions are timely has impact on use.

The End Game

Use of findings frequently occurs at the end of the evaluation, after the presentation of the final report. (Remember again that use may occur based on interim findings or based on the evaluation process.) In Section U, I spoke of the means of more effectively presenting evaluation reports. I have spoken of the evalua*tor*'s credibility earlier in this section. Evalua*tions* also may have credibility—or not. Accordingly, an *evaluation report* must be *seen as credible*. In part, report credibility is a function of the evaluation procedures employed and the extent to which they are communicated in an understandable fashion. Credible evaluation reports are more likely to lead to use.

The final area on which I comment is related to how you, as the evaluator, may assist in attaining use through active involvement in *teaching stakeholders how to use information*. Thinking about how to use evaluation information is not really part of most stakeholders' worlds. Sure, they may think about large summative decisions, but digging into the data and sorting out the implications is more difficult. At this point, I ask you to reflect on my previous comments with respect to making recommendations as part of the evaluation report. I generally "don't do" recommendations. I prefer that courses of action be determined by the stakeholders. This is where you, as the evaluator, play an important role. Your guidance in helping stakeholders understand the meaning and potential implications of findings is vital.

Getting to the point of evaluation use often requires group action. Evaluator skills in helping stakeholders to engage in the process of considering possible use is often necessary. In a study of effective evalu-

ation use that I coordinated, one of my professional colleagues noted that "School administrators, teachers, and parents . . . often do not have group process skills and decision making skills. They must be given assistance in how to read, analyze, and make decisions based on evaluation data." The effective evaluator that my colleague studied had outlined a planning and decision-making process in which the evaluator listed the specific steps sequentially, the data that were available to inform the decisions, and the specific decisions that could be made. Another professional colleague in that same study identified an evaluator who regularly encouraged stakeholders to use evaluation data by giving them information on a related task or problem. In essence, this evaluator trained stakeholders to use data in making program decisions by giving them practice with the process. The bottom line: Engage stakeholders in activities designed to focus their attention on the multiple ways that evaluation information might be used.

GUARD AGAINST MISUSE

I have talked about the importance of your actions in helping evaluation use to occur. Of equal importance is your responsibility for guarding against misuse. Misuse occurs when stakeholders modify, misstate, or inappropriately excerpt from the evaluation report. You have an important responsibility for ensuring that your evaluation is conveying the information that you intended and not being misapplied in ways not justified by your evaluation. Misuse may start with stakeholders taking your report and simply modifying sections of the report. This is inappropriate. Misuse may occur by stakeholders summarizing elements of the report in ways that are not consistent with what you stated. This is inappropriate. Misuse by stakeholders may occur when they injudiciously excerpt portions of the report consistent with their belief, but not with the tone of the report. This is inappropriate.

And so, I ask you to consider use; do all that you can to foster appropriate use. However, be alert to potential misuse.

RECAP—SECTION V

Attaining Evaluation Use

- About Use
  - Use refers to *this* program
  - Conceptual use

- Instrumental use
- Findings use
- Process use
- Evaluator Actions: The Preconditions
  - Credibility
  - Evaluator commitment to use
  - Stakeholders' prior evaluation experience
- Evaluator Actions: During the Evaluation
  - Stakeholder interest in this evaluation
  - Valued norm
  - Active participation
  - Timeliness
- Evaluator Actions: The End Game
  - Effective reports
  - Teaching use to stakeholders
- Guard against misuse

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

Consider the RUPAS evaluation that you have been working through. Who might use the evaluation findings? Assume that Family Matters will be a user and that the stakeholders at Family Matters had no prior evaluation experience. What might you have done to foster a greater inclination for the organization to use the evaluation? What might you do to engage in "teaching use," as we have discussed it, to other stakeholders (e.g., Children's Trust and others)?

Furthermore, from an evaluator's point of view, in what ways might you establish and maintain your credibility as a precondition to enhancing use? What strategies might you engage in to support others' use of evaluation within the RUPAS Program itself? How might you and your stakeholders support one another to facilitate use of evaluation findings? What actions might you take?

Further Reading

Alkin, M. C., & King, J. A. (2016). The historical development of evaluation use. *American Journal of Evaluation, 37*(4), 568–579.

My colleague Jean King and I seek to document the development of the notion of evaluation use in this article, which we trace to two important bodies of work—educational testing and measurement and knowledge utilization.

Alkin, M. C., & Taut, S. M. (2003). Unbundling evaluation use. *Studies in Educational Evaluation, 29*(1), 1–12.

This article describes the "landscape" of evaluation utilization—namely, the different kinds of utilization—with an emphasis on those that can be influenced by the evaluator.

Cousins, J. B. (Ed.). (2007). *Process use in theory, research, and practice* (New Directions for Program Evaluation No. 116). San Francisco: Jossey-Bass.

This journal volume provides an excellent introduction to process evaluation. Pay particular attention to Chapter 1, pp. 5–8, for a discussion of the historical evolution of the term "process use." Also see Chapters 3 and 7 by Jean King and Michael Patton, respectively.

Quick Reads

1. Joy Kaufman and Andrew Case on Increasing Evaluation Use through Partnership with Consumers of Services
   http://tinyurl.com/zprdby8

2. Keiko Kuji-Shikatani on Using Evaluative Thinking to Support Evaluation Use
   http://tinyurl.com/hwnakan

SECTION

W

What Are the Evaluation Standards and Codes of Behavior?

By now, you have a sense of what constitutes a good evaluation—that is, you have some understanding of the standards that might be used to *judge whether an evaluation is well done*. I would like you to reflect for a few minutes about what you have read in this book. What do you think might be the five major themes that emerge for making that judgment? Stop now. Don't turn to the next page quite yet. Think about it. Remember there is no single right answer. What overarching characteristics make for a good evaluation? *Think about it!*

JUDGING AN EVALUATION

Let us jointly consider what we have conversed about in this book. First, you undoubtedly have noticed a continued discussion on the topic of *evaluation use*. We talked about doing meaningful evaluations that can be (and hopefully would be) used by stakeholders in order to improve their programs. We highlighted the importance of identifying appropriate audiences—the stakeholders. Furthermore, we considered the questions to be examined in the evaluation and emphasized the selection of questions that stakeholders really want answers to, answers that will be useful to them. We considered the topics of report clarity and appropriate dissemination.

In Section C, we considered the issue of who does evaluations—what an evaluator looks like. In that section, I highlighted the importance of the evaluator having a use orientation. Perhaps one of the most important things leading to potential evaluation use is an attitude and a belief on the part of evaluators that it is important.

And so, we see that one potential *standard* for judging evaluations reaffirms the importance of seeking evaluation use. The goal is evaluation use. When judging the evaluation itself this means that it should have *utility*—that is, the characteristics that make it potentially useful.

What else? Clearly, an evaluation cannot be conducted if the evaluation plan and projected evaluation procedures are not realistic or practical. One can create exotic evaluation designs, which may be totally inappropriate for the context—and which can't be conducted. Evaluations have to be organizationally, politically, and financially feasible. I have addressed some of these topics within sections of this book. We jointly considered the organizational, social, and political contexts and the ways in which they enable or constrain the evaluation. We also looked at whether the questions being considered were evaluable, taking into consideration these organizational, social, and political elements as well as cost considerations—that is, we were trying to ensure that the evaluation to be conducted was *feasible*.

In several places throughout the book we talked about ethical issues. We considered ethics from the point of view of the evaluator in terms of disclosing potential conflicts of interest. I also stressed the necessity of the evaluator not being compromised in the presentation of evaluation reports. The evaluator also has to deal with stakeholders in a manner that respects their rights and their points of view. Furthermore, at various junctures in this book I indicated the need to go into the community and try to understand "where people are coming from." I noted the need to get the broad stakeholder views that could be interjected into the discussion with primary stakeholders. I indicated

the need for respecting the rights of human subjects and respectfully dealing with people. I commented on the concern for the evaluator to be aware of the legal constraints surrounding the evaluation, including the kinds of mandated reporting that would be required. Let's give a name to this assortment of legal, ethical, and personal interaction elements and call it *propriety*—the condition of being proper, fitting, or suitable.

Of course, what is an evaluation if it is not technically correct—technically accurate? We have talked about evaluation as being a systematic procedure for the collection, analysis, and valuing of information. And indeed, *accuracy* needs to be considered as one of the standards for judging a good evaluation. This entire book deals with conducting a technically correct and adequate evaluation. I maintain that the first element of accuracy is getting it right—asking the right questions. We certainly have explored that issue in many, many sections. Identifying the appropriate information sources is another element of accuracy. The development of an evaluation plan with defensible information sources keyed to the evaluation questions is yet another step. Appropriate collection and analysis of data—valid and reliable data—are likewise important. Finally, reaching informed conclusions is an essential element of accuracy.

THE PROGRAM EVALUATION STANDARDS

I CHEATED. The above discussion, while absolutely true and reflecting the issues discussed within this book, identifies four of the five major concepts in the book *The Program Evaluation Standards,* which was produced by the Joint Committee on Standards for Educational Evaluation (JCSEE). A few words of description are in order here.

The JCSEE was initially formed in the late 1970s. The committee consisted of representatives from 16 major professional associations. Their work engaged hundreds of persons in deliberations, writing, field testing, and the like, in order to develop agreed-upon standards for the conduct of evaluation. The first publication of *The Program Evaluation Standards* was in 1981, with subsequent editions in 1994 and 2011. The embracing concepts of *utility, feasibility, propriety,* and *accuracy* have formed the basis for the *Standards* since its initial publication.

In the 2011 edition, a fifth concept was added: *evaluation accountability.* In essence, this refers to the importance of systematic reflection and examination of the extent to which an evaluation is accountable. I tend to think of the accountability notion more broadly. The standard on accountability refers to a responsibility to stakeholders to obtain a systematic review of their evaluation in order to provide assurance that the evaluation was done well. I believe that the activity called metaevaluation, which forms the basis of the accountability concept, is only one part of an evaluator's actions to evaluate his or her own efforts in order to improve. I discuss the need for learning more about evaluation and for improving evaluator skills in Section Z.

Within each of these broad standards, the JCSEE has derived from three to eight specific standard statements. The *Program Evaluation Standards* summary is presented in Table W.1. The standard statements are displayed by category (e.g., "Utility")—that is, each category pertains to a particular type of issue.

To clarify, the standard statements as a group provide an indication as to how to attain utility in evaluations. However, standard statements do not correspond directly with the particular functions involved in conducting an evaluation. Thus, particular standard statements within a category are applicable to the various activities involved in conducting an evaluation—the sections within this book. Some standard statements might have applicability to only a limited number of sections. Others might be very broad and relevant to many sections. For example, standard statement U1, "evaluator credibility"; statement U2, "attention to stakeholders"; and statement U8, "concern for consequences and influence" are relevant to the way in which we have discussed evaluation in many sections of this book. They are essential elements in attaining use through having use as a focus by the evalu-

ator, who has credibility and seeks the involvement of stakeholders. Standard statement U3, "negotiated purposes," is also relevant to many aspects of the evaluation, but perhaps is not as overarching as the prior three statements. The idea of understanding stakeholders' purposes extends to the development of the logic model, the focusing on evaluable questions, developing the design, analyzing results, and reporting. Other evaluation standards are more specifically applicable to particular evaluation functions. It would be helpful as you read the specific standard statements to reflect on the ways in which each of them has been dealt with in this book.

TABLE W.1. The Program Evaluation Standards

Utility standard statements	
U1 Evaluator Credibility	Evaluations should be conducted by qualified people who establish and maintain credibility in the evaluation context.
U2 Attention to Stakeholders	Evaluations should devote attention to the full range of individuals and groups invested in the program and affected by its evaluation.
U3 Negotiated Purposes	Evaluation purposes should be identified and continually negotiated based on the needs of stakeholders.
U4 Explicit Values	Evaluations should clarify and specify the individual and cultural values underpinning purposes, processes, and judgments.
U5 Relevant Information	Evaluation information should serve the identified and emergent needs of stakeholders.
U6 Meaningful Processes and Products	Evaluations should construct activities, descriptions, and judgments in ways that encourage participants to rediscover, reinterpret, or revise their understandings and behaviors.
U7 Timely and Appropriate Communicating and Reporting	Evaluations should attend to the continuing information needs of their multiple audiences.
U8 Concern for Consequences and Influence	Evaluations should promote responsible and adaptive use while guarding against unintended negative consequences and misuse.

(continued)

TABLE W.1. *(continued)*

Feasibility standard statements	
F1 Project Management	Evaluations should use effective project management strategies.
F2 Practical Procedures	Evaluation procedures should be practical and responsive to the way the program operates.
F3 Contextual Viability	Evaluations should recognize, monitor, and balance the cultural and political interests and needs of individuals and groups.
F4 Resource Use	Evaluations should use resources efficiently and effectively.
Propriety standard statements	
P1 Responsive and Inclusive Orientation	Evaluations should be responsive to stakeholders and their communities.
P2 Formal Agreements	Evaluation agreements should be negotiated to make obligations explicit and take into account the needs, expectations, and cultural contexts of clients and other stakeholders.
P3 Human Rights and Respect	Evaluations should be designed and conducted to protect human and legal rights and maintain the dignity of participants and other stakeholders.
P4 Clarity and Fairness	Evaluations should be understandable and fair in addressing stakeholder needs and purposes.
P5 Transparency and Disclosure	Evaluations should provide complete descriptions of findings, limitations, and conclusions to all stakeholders, unless doing so would violate legal and propriety obligations.
P6 Conflicts of Interests	Evaluations should openly and honestly identify and address real or perceived conflicts of interests that may compromise the evaluation.
P7 Fiscal Responsibility	Evaluations should account for all expended resources and comply with sound fiscal procedures and processes.

(continued)

TABLE W.1. *(continued)*

Accuracy standard statements	
A1 Justified Conclusions and Decisions	Evaluation conclusions and decisions should be explicitly justified in the cultures and contexts where they have consequences.
A2 Valid Information	Evaluation information should serve the intended purposes and support valid interpretations.
A3 Reliable Information	Evaluation procedures should yield sufficiently dependable and consistent information for the intended uses.
A4 Explicit Program and Context Descriptions	Evaluations should document programs and their contexts with appropriate detail and scope for the evaluation purposes.
A5 Information Management	Evaluations should employ systematic information collection, review, verification, and storage methods.
A6 Sound Designs and Analyses	Evaluations should employ technically adequate designs and analyses that are appropriate for the evaluation purposes.
A7 Explicit Evaluation Reasoning	Evaluation reasoning leading from information and analyses to findings, interpretations, conclusions, and judgments should be clearly and completely documented.
A8 Communication and Reporting	Evaluation communications should have adequate scope and guard against misconceptions, biases, distortions, and errors.

(continued)

TABLE W.1. *(continued)*

Evaluation accountability standards	
E1 Evaluation Documentation	Evaluations should fully document their negotiated purposes and implemented designs, procedures, data, and outcomes.
E2 Internal Metaevaluation	Evaluators should use these and other applicable standards to examine the accountability of the evaluation design, procedures employed, information collected, and outcomes.
E3 External Metaevaluation	Program evaluation sponsors, clients, evaluators, and other stakeholders should encourage the conduct of external metaevaluations using these and other applicable standards.

Note. From *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users, Third Edition*, by Donald B. Yarbrough, Lyn M. Shulha, Rodney K. Hopson, and Flora A. Caruthers. Copyright © 2011 SAGE Publications, Inc. Reprinted by permission.

AMERICAN EVALUATION ASSOCIATION GUIDING PRINCIPLES FOR EVALUATORS

There is another set of guidelines that is used by evaluators. The American Evaluation Association (AEA) created a task force to consider guiding principles for evaluators. It is important for us to differentiate between the *Program Evaluation Standards* and the *AEA Guiding Principles for Evaluators*. As you noticed in the prior discussion, the *Evaluation Standards* focus on the characteristics of an evalua*tion* that make it a good evaluation. On the other hand, the *Guiding Principles for Evaluators* have as a primary focus the actions that the evalua*tor* should (or should not) take.

The five major categories of the *Guiding Principles for Evaluators* published by the AEA are:

1. Systematic inquiry
2. Competence
3. Integrity/honesty
4. Respect for people
5. Responsibilities for general and public welfare

The first of the five guidelines focuses on *systematic inquiry*. To adhere to this guideline, it is necessary to understand the purposes of the evaluation, the evaluation questions (often derived from the logic model), appropriate design, analysis, and reporting. To which I would add the importance of careful attention to encouraging potential use throughout the process. In essence, systematic inquiry is what we have been talking about in this book and is further reflected in the *Standards* established by the JCSEE.

Now we turn to the three guidelines that are related to personal characteristics and the behavior of the evaluator. These are *competence, integrity and honesty,* and *respect for people*. For the *competence* guideline you, as the evaluator, should ensure that you and the evaluation team collectively possess abilities, skills, and experience appropriate to the evaluation, and commit yourself to maintaining and improving those competencies as needed. Furthermore, the evaluation team should collectively demonstrate cultural competence and then employ evaluation strategies appropriate to the culturally different groups.

A third guideline dictates that evaluators should display *integrity and honesty* in their own behavior in such things as negotiating honestly with clients and relevant stakeholders and avoiding apparent conflicts of interest. Honesty and integrity are also displayed by conducting and reporting evaluation results in an honest unbiased manner to avoid providing misleading evaluation information and to help prevent misuse of evaluation results.

Evaluators should also show *respect for people*. This guideline refers to respecting the dignity and self-worth of evaluation stakeholders. The basic concern of this guideline is that the evaluator adhere to appropriate professional ethics, standards, and regulations regarding confidentiality, informed consent, and potential risks or harms to participants. It is also important that you, as the evaluator, seek to understand contextual elements so you can take into account differences among stakeholders such as culture, religion, disability, age, sexual orientation, and ethnicity.

The fifth guideline, *responsibilities for general and public welfare,* offers the perspective that the evaluator must take into account the diversity of general and public interests and values including relevant perspectives and interests of the full range of stakeholders. In doing so, you should consider the implications of your evaluation on matters of public interest and the welfare of society as a whole.

RECAP—SECTION W

Evaluation Standards and Codes of Behavior

- The Joint Committee Program Evaluation Standards
  - Utility
  - Feasibility
  - Propriety
  - Accuracy
  - Evaluation accountability
- Individual Standard Statements May Apply to Multiple Evaluation Functions
- American Evaluation Association Guiding Principles for Evaluators
  - Systematic inquiry
  - Competence
  - Integrity/honesty
  - Respect for people
  - Responsibilities for general and public welfare

GAINING ADDITIONAL UNDERSTANDING

Evaluation of RUPAS

The *Program Evaluation Standards* and the *Guiding Principles for Evaluators* are helpful when trying to judge the quality of an evaluation and the evaluator's competence. Reflect on what you have learned about the *Standards* and codes of behavior. Again, consider the RUPAS case and think about some of the difficulties that might have prevented adherence to the *Evaluation Standards*. For instance, which of the evaluation activities that you considered pursuing seemed to be in compliance with the *Standards*? Which may not have been appropriate? Consider, also, ways in which you, as the evaluator, might not have been sufficiently mindful of the *Guiding Principles*.

Further Reading

Morris, M. (2011). The good, the bad, and the evaluator: 25 years of AJE ethics. *American Journal of Evaluation, 32*(1), 134–151.

This article is organized by the five *Guiding Principles for Evaluators* and highlights key insights about ethical evaluation practice in each.

Morris, M., & Clark, B. (2013). You want me to do what?: Evaluators and the pressure to misrepresent findings. *American Journal of Evaluation, 34*(1), 57–70.

This article reports on results from a survey study of over 2,500 American Evaluation Association members concerning the pressure to misrepresent evaluation results and the various ways in which this dilemma was resolved.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: SAGE.

Approved by the Joint Committee on Standards for Educational Evaluation, an American National Standards Institute member, this book gives the full description of the standard statements with clarifying examples.

Quick Reads

1. Kathy Bolland on Ethics in Evaluation and in Social Professions
   http://tinyurl.com/o2emdc5
2. Gail Vallance Barrington on Living Your Ethics
   http://tinyurl.com/zylow6y
3. Katye Perry on Teaching Ethics in Evaluation through Case Examples
   http://tinyurl.com/jdgelc8

SECTION

X

How Are Costs Analyzed?

Cost analyses are only rarely included in evaluations. In many ways this is unfortunate because understanding the costs of a program relative to its benefits or effectiveness is important. But cost analyses of various types can be very difficult to do—indeed, you might ask: Is a cost analysis of a program as part of an evaluation worth the cost?

Let me provide an example to simply show the difficulties of performing an analysis of *just* the costs of a program and not its performance. To do this and stick to the familiar, I select as an example a product evaluation instead of a more complex program evaluation. Assume, for example, that you are thinking of buying a home theater system. The first question you might ask is what such an entity costs: What is the *cost*? This is a seemingly simple question. *It is not!* You might indicate that the answer is to be found by simply checking a few large appliance stores or Amazon and seeing what price they are charging. However, that is not the real cost. Determining the real cost of a program (in this case, a product) is a very difficult enterprise. Think, for example, of the cost of traveling to the stores, or of setting the system up at home, or the cost of express shipping from Amazon. There are many other personal costs as well. However, I do not get into that now, but defer the discussion of "costing" to the end of this section. At any rate, you perhaps now have a minimal understanding of why cost analyses are so difficult.

COST-EFFECTIVENESS ANALYSIS

One kind of cost analysis question is "Where can we get the best deal?" Imagine that you are considering two stores: Target and a local appliance chain. Each has a home theater system of interest to you. The old story of "comparing apples and oranges" may come into play here. There are two ways in which a comparison might be done. The simplest is that if they both carry the same system and it is the one that you desire, then you can compare costs. To make it simpler, think of the price the store charges as the total cost—but keep in mind this will be shown to be an incomplete accounting of the real cost. At this point, and with no added considerations, you might choose the cheaper system as your most cost-effective option.

On the other hand, if each store has a different system but at the same price, then you can compare them on the capability—such as the most volume without distortion on your favorite DVD. You would have an easy way to choose. But because you may have several qualities of a system, such as volume and number and types of ports for accessories, a decision will require a matter of personal judgment in *weighting* the relative importance of the various quality or outcome measures.

In either case, considering the same product at different prices, or different products at the same price, can benefit from cost-effectiveness analysis. You are comparing two items on the basis of their costs and their effectiveness (in the above example, effectiveness is their capability).

Let's apply this framework to a health program. Performing a cost-effectiveness analysis of a drug treatment program, for example, requires that you have a competing program. Moreover, if that competing program is of the *same cost,* then you can *compare* the accomplishments or *outcomes* of the programs. Which program keeps addicts off of drugs for the longest period of time may be the central outcome—or some other appropriate outcome measure.

An alternative condition that allows the conduct of a cost-effectiveness evaluation is to find a competing drug treatment program that perhaps uses different methods, but has the *same specific goal,* and the two programs *attain that goal equally*. For example, 75% or more of the participants stay off of drugs for 3 months. Assume that two competing programs meet that goal. In that instance, you can *compare their costs*. You can see from this example how difficult it is to find two programs with such comparability in order for a cost-effectiveness analysis to be completed.

As I have noted, the major impediment to the conduct of cost-effectiveness evaluation is the necessity of restricting the analysis to a

single benefit measure. Cost-effectiveness can be done only if two programs have equal costs—then differences in outcomes can be examined. Or if a single outcome (benefit) is the same in each program, then costs can be compared. This is, indeed, a major dilemma because most social programs have multiple desired outcomes.

COST–BENEFIT ANALYSIS

Another kind of cost analysis procedure, even more sophisticated and complex, is referred to as a cost–benefit analysis. This analysis shows the relationship between the costs of a program and the benefits that will be attained—that is, if a program costs $50,000, will the value that will be derived equal or exceed that dollar amount? This, of course, requires that the benefit measure be converted to dollars. So in the simple home theater example, if you are contemplating the purchase, you will need to determine the actual dollar benefits that might be derived from having a home theater system. Will it provide learning experiences that better prepare you to obtain particular jobs? What will be your additional lifetime income obtained above your current expectations? Will it keep your teenage children off the streets and out of trouble? What kind of possible trouble and how might that impact their future benefits (in dollars)? Wow!!

Consider, also, another program example. Imagine that 85 people complete a drug treatment program. How might you, as the evaluator, determine in dollar terms the various benefits associated with that program completion? In this program case, you might be able to examine the societal benefits but not the personal benefits. The government has established a program. Is it a worthwhile expenditure for society? Do the benefits exceed the costs? To do this, you will need to investigate what research evidence says about the cost to society of an individual who returns to using drugs. What about police costs? Social service costs? But there are also personal benefits. An individual who has successfully completed the program may be in a better position to obtain employment, get off of welfare, and contribute to society. In small programs, such as in this example, the extreme difficulty of doing these calculations makes cost–benefit analysis not cost beneficial. (That's a pun.) The cost of aggregating data on benefits in dollar terms might be a very large percentage of the cost of running the program itself. The cost of performing that analysis will surely exceed the derived benefits to the program.

However, cost–benefit analyses have been employed productively on very large-scale social programs. Economists, for example, have calculated the costs and benefits of building a dam, and in doing so, they

have looked at costs including the social displacement of individuals; the benefits from the electricity produced by such a dam; and many, many other costs and benefits.

As we have seen, the dilemma with cost-effectiveness analysis is that we are restricted to a single outcome measure. If we are comparing two programs, they must have the same desired outcome, measured in the same way, and we may only do the comparison based on a single outcome. The dilemma with cost–benefit analysis is that somehow we must find a way to convert all outcome data (the benefits) *to dollar amounts*.

COST–UTILITY

Some economists talk about a potential way out of these dilemmas. They suggest modifying cost-effectiveness procedures by using subjective judgments as a measure of outcomes. The term employed is "utility," meaning the degree of happiness, or the degree of benefit that one *perceives* has been obtained. Note the word "perceives." The utility notion comes from the philosophical concept of utilitarianism, an approach to ethics generally attributed to Jeremy Bentham and expanded by John Stuart Mill. The way in which this concept might be used in a cost analysis framework is in the narrower economic sense of perceived usefulness.

The concept of cost–utility is deceivingly simple. Indeed, it is instinctual. You look at the cost of several competing alternatives and then you look at the utility rating or happiness that is to be derived from each alternative. One TV costs $1,000 and another costs $1,500. Will you get more happiness from purchasing the first or the second one? Suppose you compare them more formally. Imagine that you assign utility ratings to each of the TV sets, but, of course, there must be ground rules to guide your utility ratings. Assume that you rate each on a scale of 0–10. Furthermore, the scale ratings need to be of equal value—that is, the data are like interval data (see Section R). Actually, given that the data are grounded at 0, they are what is called "ratio data," the data type that we previously alluded to. If you rate one TV as 4, then a rating of 8 is considered twice as good. Having a utility rating for each, you then can make a calculation of cost per utility. This is demonstrated in Table X.1.

So we see that TV set 2—which costs $500 more and has twice the utility of TV set 1—provides twice the happiness. When we divide "cost" by "utility" it is attained at a cost of $200 per utility unit (utile), rather than $250—that is, it costs less per unit of utility or happiness.

Now, let me discuss how a cost–utility analysis can be performed at the program level, but first a few thoughts about the word *utility*. Economists who talk about the notion of utility fret about the near

TABLE X.1. The Concept of Cost–Utility

	Cost (C)	Utility (U) (10-point scale)	C/U ratio
TV Set 1	$1,000	4	1,000/4 = 250
TV Set 2	$1,500	8	1,500/8 = 200

impossibility of obtaining an accurate utility measure. I find that to be a relative nonissue for our purposes. For the most part, they are concerned about large-scale, multisite programs and thus about generalizability issues. They question whether any sample population can truly reflect real societal utilities. I have approached evaluation in this book as reasonably site specific and as attempting to meet the needs of local programs. Thus, since generalizability is not an issue, the perceived utility of local stakeholders should suffice quite nicely. We want to consider what those whose program it is, and who live with the program, consider to be of value—that is, to have high utility.

Therefore, we wonder how you, as the evaluator, might go about doing a cost–utility study in a local context given the context-sensitive participatory mode that I have presented in this book. Let me deal here only with the discussion of the utility component of the cost–utility analysis. Naturally, costs will have to be determined in some systematic manner, as I discuss later in this section.

First, let me make it clear that the *utility measures* need to be *determined prior* to any data *collection* or *analysis*. We do not want the results, as subsequently noted by stakeholders, to bias perceptions about the degree of importance of any of the outcome areas. The method that I propose is a more sophisticated version of that used as an example in Section A. In that instance, where I was deciding on a house to buy, I delineated the aspects of a house considered to be important (think of these as outcome measures) and then indicated the relative importance of each aspect. In essence, I was asking the degree of happiness (utility) that might be ascribed to each attribute—an extra bathroom, for example. In the case of a real evaluation, we would have more accurately defined outcome measures.

SINGLE OUTCOME

So let me describe a potential procedure for determining utility where there are several alternative programs, each of which has the same *sin-*

gle measurable outcome. Let us say, for example, a test score. This discussion may be somewhat difficult so I suggest you look at the relevant table as I discuss this material. The question facing us is how much decision makers are willing to pay per increment of test score increase. For example, if one program receives an average score of 500, how much greater happiness is derived with the achievement of a score of 510? Would they be 10% happier or 20% or 50% happier? In essence, we need to establish outcome score rankings of satisfaction (or happiness) for scores that might be obtained. Again, I suggest a 0–10 rating. The stakeholder rankings for the different scores might look something like what is represented in Table X.2.

Assume that this chart recognizes an expectation of a score of 500 and is assigned a ranking of 5. A score of 490 expresses a degree of dissatisfaction and is assigned a score of 2. A score of 480 or below is considered worthless—that is, the program is a failure. On the positive side, a 10-point increment above 500 seems like a plausible and achievable goal, and a score of 510 might warrant a jump to a utility rating of 7. Scores above that have only incrementally larger utility ratings. In this instance of a single ranking, this ranking is, in essence, the utility measure.

Then, dividing the cost of each of the competing programs by its ranking provides a cost–utility ratio. When using the rankings in Table X.2, if one program has a cost of $800 and a rating of 8 (which is a score of 520), then its cost per utility rating ratio is 100 ($800/8). A competing program with a rating of 5 (score of 500) and a cost of $750 would have a cost per utility rating of 150 ($750/5). The former program has a lower cost–utility ranking—it is less costly per utile—and would be judged best.

TABLE X.2. Stakeholder Rankings of Utility

Score	Outcome attainment ranking (R)
480 or below	0
490	2
500	5
510	7
520	8
530	9
540 or above	10

MULTIPLE OUTCOMES

A program having multiple outcomes builds upon this procedure. We now know how to calculate the utility for differences in an outcome. The same procedure described above can be used for each of the program's outcome areas. But now the question arises: How do we rank the differences in outcome attainment? We discussed this in the previous discussion of single measures. So we see in Table X.3 that we have multiple outcomes. Each of these outcome areas can have an importance measure attached to it. Again, it is necessary to stress to the primary stakeholders that they are to do the ranking for each outcome on a scale of 0–10 where 0 is considered as providing no contribution to happiness, 5 being average, and 10 representing the highest possible contribution. Furthermore, it is important that the scales be treated as if they have interval properties—that is, a score of 6 is viewed as having twice the utility of a score of 3.

Now let us look further at Table X.3. We have two programs that I have creatively called Program 1 and Program 2. For each of these programs there are two outcome measures: reading score and student satisfaction. Through a process depicted in Table X.1, we have determined with stakeholders that reading score is of more importance (I)—it has a larger utility weighting—than student satisfaction. Thus, reading score received a relative importance weighting of 6 and student satisfaction an importance weighting of 4—that is, in considering these two outcome measures, the reading score is 50% more important than student satisfaction (6 compared to 4, respectively). These weightings are listed in Column 2 (marked as "I") for each program. Furthermore, outcome scores were converted into an outcome ranking for each measure based on the happiness value of each outcome (in a manner similar to Table X.2). In this case, we developed an outcome attainment ranking table showing how much satisfaction (or happiness) is attained from different levels of achievement for each of the two outcome measures. These score rankings are shown in Column 1.

The *total measure utility ranking* (U), as shown in Column 3, is simply obtained by multiplying the "R" and the "I" values. The *total program utility ranking* is obtained by adding the score for reading and the score for student satisfaction for each program. Thus, in this instance, Program 1 has a total program utility of 68 and Program 2 has a total program utility of 76. Dividing the cost per participant for each program by the total program utility score provides us with a cost–utility (C/U) ratio. Program 1 has a C/U ratio of 13.23 and Program 2 has a C/U ratio of 12.13. Program 2, then, is the favored program because its cost of $12.13 per utile (unit of utility) is less.

TABLE X.3. Establishing a Cost–Utility Ratio

Outcome measures	R* Outcome score ranking (1)	I Relative importance of that measure (2)	U Total measure utility ranking (3)	Total program utility ranking (4)	Cost per participant (5)	C/R ratio (6)
Program 1						
A (e.g., student satisfaction)	8	4	32	68	$900	$900/68 = $13.23
B (e.g., reading score)	6	6	36			
Program 2						
A (e.g., student satisfaction)	7	4	28	76	$960	$960/76 = $12.13
B (e.g., reading score)	8	6	48			

*As described in the previous example (Table X.2).

AND NOW TO COSTS

What are the program costs? Simple! Can't we just simply look at the budget for the program? No! And no!

The real cost for the home theater system that we talked about earlier must include your time in going to the store, the transportation cost, and the installation cost once you get it home. For programs that are to be evaluated, one can only imagine all of the data that need to be aggregated in order to compute a real cost. Calculating costs is a very complex endeavor. Some authors have provided quite detailed descriptions about how to identify and determine the costs for the *ingredients* of programs. We provide a reference in the Further Reading for this section to those inclined.

Think for a moment. A new program is being introduced. What is required in order for the program to be able to operate? The budget provides some insight but not enough. Of course, there are standard budget items such as a listing of the personnel who will be required to run the program. This may include staff employed to directly deal with clients in operationalizing the program, but it will also need to include administrative personnel.

The program may require various materials or equipment necessary to operate the program—computers, for example. What supplies are necessary? Paper, pencils, or pens? But wait, there are also additional administratively related costs. Perhaps clerical help is required for ordering supplies or following up with clients. What about custodial staff for maintaining or renovating the facility? Now what about consultants? Possibly consultants are needed for equipment maintenance. Furthermore, will staff training be necessary, and if so, will consultants be needed for this training?

"But wait," you might say, "What about the space we already have?" Furthermore, there are no new administrative personnel in the larger organization who will be providing various administrative backup services, and the facility is already being maintained for the other programs currently taking place. So how do you, as the evaluator, handle this? The answer: It depends. If two alternative programs are being implemented at the same site using the same amount of facilities and comparable staff, then many of the costs that one might want to determine are in essence the same for both programs. In that instance, it might suffice to simply look at marginal (or additional) costs that are associated with each individual program—that is, one program might need additional equipment while another might need additional aides. There might be unique travel expenses associated with one program and not with the other. By looking at the marginal costs, we would

not determine the actual full cost of each program, but would gain an understanding of the relative cost differences that could be used in a cost analysis (cost-effectiveness analysis, cost–benefit analysis, or cost–utility analysis). If the programs have substantially different cost bases, then it may not suffice to look only at the marginal costs. In that case, full costing would be required.

HOW TO DETERMINE COST

One influential evaluator has suggested a procedure of identifying all of the ingredients of a program that have costs associated with them. I pretty much have alluded to the categories in the previous discussion: personnel, space, equipment, supplies, and services (including consultants).

Perhaps the most straightforward costing takes place in regard to *personnel*. Of course there is a listed salary for each individual directly involved. Associated with the salary are employee benefits that also need to be determined. The two alternative programs being compared in the cost analysis may have staff employed who have slightly different salaries based on, perhaps, their seniority, but if there are no specific requirements for years of service necessary for participation in the program, then it is important not to allow the biasing effect of particular salaries. Thus, it is appropriate to determine a typical salary for that employee category for each program.

A particularly vexing issue is how to deal with personnel who have only a portion of their time allocated to participating in the program. Thus, an aide might be assisting in two different programs. Clearly, one needs to determine the portion of the time to be charged to each program. This is also an issue with administrative personnel who may be responsible for a whole organization and who only peripherally supervise the program in question.

Let us next consider *facilities/space*. Will it need to be rented? At its simplest we just include as the cost for space what the actual rental price is for the facility of the type required. Or if a space is already available, then we include the cost of renting a comparable space. This is not without its difficulties. If the typical available space that might be required for operating the program would require renovation and restructuring in order to meet the needs of the program, then these particular renovation costs would also need to be included. While this remodeling or restructuring might take place in the first year of the program's operation, you will need to amortize the cost of this remodeling over multiple years.

Another area for which costs need to be determined is *equipment*. Some large equipment might feasibly be rented or leased. However, many small purchases, such as a computer or a printer, are more appropriately purchased, and (I'm sure you hear it coming) this purchase price needs to be spread over the expected life of that equipment.

As alluded to earlier in this section, the program might require *consultants* of one type or another. Some consultant services might be required throughout the program. For instance, there might be the need for an ongoing computer consultant. But in the case where program staff need to be trained on-site or participate in a more general off-site training workshop, such training might be required only in the first year of the program, or perhaps every 2 or 3 years. Thus, this cost cannot be charged in a single year. (And . . . you know what comes next about how to treat this cost.)

Finally, there might be *travel costs*. Do these occur each year or are they primarily in the first year? The answer will help you to determine how these costs are handled. Furthermore, there certainly are costs associated with maintaining the facility and providing utilities (heat, lighting, telephone, etc.). What about supplies: paper, unique program materials, printing, food to be provided to participants, publicity materials, brochures?

There are also costs to entities other than the program. For example, *costs incurred by participants*—perhaps purchases that they need to make, or transportation. Time is a cost. For example, more often than you wish, you might have made the statement, "That meeting just cost me 2 hours." The inference is that time has value. There are other things that might be done using that time. Different programs may require different amounts of client time. Whether one wants to consider this time as a cost is perhaps debatable. It depends on whether you are considering the costs simply from the perspective of the program, or not—that is, if you are considering your cost analysis only from the point of view of the program, you might not include participant time as a cost, but if you take a more societal benefit perspective, then you will include these time costs. Another cost related to time that might be included (depending on the perspective one takes) is volunteer time. If individuals are volunteering to assist in a program, they are saving personnel costs. Moreover, their volunteer time might well have been coaxed into participating in another program.

The costing elements that I have been discussing and an efficient means for examining them are presented in a book by Levin and McEwan listed in the Further Reading for this section. Full understanding of the cost issue requires greater detail than I am able to provide here.

RECAP—SECTION X

Cost-Analysis Procedures

- Cost-Effectiveness Analyses (Multiple Programs)
  - Determine costs
  - Determine (single measure) outcomes
  - Same costs—compare on outcomes; highest outcome is best
  - If same outcome—compare on costs; lowest cost-effectiveness is best
- Cost–Benefit Analysis (Single or Multiple Programs)
  - Determine costs of single or multiple programs
  - Determine outcomes of single or multiple programs
  - Convert outcomes to benefits (in dollars) of each
  - Calculate ratio of costs to benefits of each
  - Lowest cost–benefit ratio is best
- Cost–Utility Analysis Multiple Programs (Single-Outcome Measures)
  - Determine costs of each program
  - Establish outcome score ranking for each level of outcome
  - Determine outcome score for each program
  - Convert outcomes to a ranking score
  - Calculate ratio of cost to rankings—lowest ratio is best
- Cost–Utility Analysis Multiple Programs (Multiple-Outcome Measures)
  - Determine costs
  - Determine a ranking score for each outcome measure
  - Determine outcome scores for each measure of each program
  - Assign this ranking score to each attained outcome measure (R)
  - Determine a relative importance score for each outcome dimension (I)
  - Multiply R $\times$ I to determine a measure of utility (U)
  - Add utility ranking of each outcome within a program to determine *total* program utility
  - Divide total program utility by cost to obtain a cost–utility ratio for each program
  - Lowest ratio is best

GAINING ADDITIONAL UNDERSTANDING

Thought Questions

You might have noticed that cost analysis is a specific family of quantitative analysis. As such, the issues and challenges that were raised in Section R still apply here. How might the ways in which outcomes are defined influence analysis of costs? Issues of stakeholder participation and valuing are salient here as well. Can you think of some reasons why that is the case?

Further Reading

Levin, H. (2005). Cost–benefit analysis. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 86–90). Thousand Oaks, CA: SAGE.

Henry Levin has produced an easily readable description of cost–benefit analyses in this encyclopedia entry.

Levin, H. M., & McEwan, P. J. (2003). Cost-effectiveness analysis as an evaluation tool. In T. Kellaghan & D. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 125–152). Norwell, MA: Kluwer Academic.

This is the definitive discussion on cost-effectiveness, cost–benefit, cost–utility, and cost "ingredients."

Quick Reads

1. Brian Yates on Doing Cost-Inclusive Evaluation, Part I: Measuring Costs
   http://tinyurl.com/hu927v7
2. Brian Yates on Doing Cost-Inclusive Evaluation, Part II: Measuring Monetary Benefits
   http://tinyurl.com/zcqtj5q
3. Brian Yates on Doing Cost-Inclusive Evaluation, Part III: Cost–Benefit Analysis
   http://tinyurl.com/zukarcz
4. Brian Yates on Doing Cost-Inclusive Evaluation, Part IV: Cost-Effectiveness Analysis and Cost–Utility Analysis
   http://tinyurl.com/jp58wnn

SECTION

Y

What Is the "Theme" of This Book?

In the various sections of this book I have prescribed how I think you should engage in the practice of conducting an evaluation. Many of the activities that I have mentioned are common to what you will find in other evaluation textbooks. But, as we discussed earlier, individuals have their own cultures and sets of beliefs that shape how they think. Let me tell you a bit about some of the beliefs that have shaped my views on evaluation, and then I will talk about the way that they have impacted my unique prescriptive theory—the theme of this book.

First, let me comment about the term *prescriptive theory*. You may tend to think of a theory as you have come to know them in the physical and biological sciences—things like Einstein's theory of relativity or Newton's law. Unfortunately, we do not have theories in evaluation that look like that. There simply is not a sufficient body of knowledge about what happens in an evaluation to be able to predict with certainty what would happen when an evaluation is employed in a particular way. These are what I would refer to as "descriptive theories" of evaluation. They do not exist.

This difficulty in potentially constructing a descriptive theory of evaluation derives from the essential difference between research and evaluation, which we talked about in Section A. We noted that research seeks conclusions and has as a goal adding to the body of knowledge. Evaluation, on the other hand, seeks to assist in making decisions or otherwise improving programs. This requires unique attention to the particular context of the program being evaluated. In order to be used,

evaluations must reflect the needs, concerns, and questions of importance to the program.

In part, this inability to construct a descriptive theory of evaluation, or multiple descriptive theories, is attributable to the fact that there are an enormous variety of contexts in which evaluation might take place—different communities, different kinds of programs, different user groups, and so on. A descriptive theory would need to consider all of these contingencies.

HISTORICAL PERSPECTIVE

So in the absence of descriptive theories, what guides the conduct of evaluations? To answer that question we must take a brief step back in time. An important era in the history of evaluation is the 1960s during the Lyndon Johnson administration when the so-called Great Society programs were implemented. Many (if not most) of these programs had requirements that evaluations be conducted. First attempts at conducting such evaluations were not promising. There was a notable lack of success in evaluating these programs. This primarily was attributable to inappropriate attempts by researchers to conduct evaluation as though it was research—but *you* know better. In response to the criticisms of the evaluations that ensued, major evaluation writers attempted to elaborate the way in which evaluations could be conducted more appropriately. In essence, major evaluation writers "prescribed" the way in which *they* felt that evaluation should be conducted. These then are what I call *prescriptive theories.*

MY PRESCRIPTIVE THEORY

How do I describe my prescriptive theory? The core principle of my approach to evaluation is concerned with *attaining evaluation use*. I care about doing something worthwhile, about having impact. I don't want to engage in some *pro forma* act of conducting technically wonderful evaluations that sit on a shelf, unused. Let me then say that the *goal of evaluation* from my perspective *is evaluation use*. This goal of evaluation use is highlighted in Section B where I discuss the issue of why do evaluations. This is neatly summarized by the paragraph at the end of Section B: "We do professional evaluation to add to an organization's ability to learn about its program, to provide a basis for judging the accountability of programs, to allow better decisions to be made (currently or in the future), and to further an organization's capacity

to continue to benefit from evaluation." I believe that major concern and attention to attaining evaluation use will foster these objectives. Discussion of evaluation use as the major goal is also readily apparent in most of the rest of this book.

I have been influenced by the writings of Robin Miller—particularly her 2010 paper found as a reference at the end of this section. She points out that what distinguishes prescriptive theories, one from another, are the particular unique ways that they go about attempting to attain their goal. She calls this a theory's "theoretical signature." (Sorry to bother you with this technical jargon but I really like the term.) This set of attributes (theoretical signature) goes beyond the actions that are typically employed in all or most evaluations.

THE RESEARCH ORIGINS OF MY THINKING

So what is the theoretical signature (the identifying elements) of context-sensitive evaluation—the evaluation approach that I espouse? Where does it come from? I have spent a good portion of my career studying evaluation use and, in particular, the factors that tend to be associated with use occurring. Many others have conducted research on this topic that added to my understanding about what kinds of activities and attributes are not likely to add to the use of evaluation. This topic—evaluation use—is a very important part of my thinking. What does this extensive research tell us?

I choose to think of these factors within four categories—the evalua*tor* category, the *user* category, the *organizational/cultural* category, and the evalua*tion* category. The listing of factors within each of these subareas of the total context is found in Appendix A. You will note that I have examined the identified evaluation use factors in two columns—where the factor as identified was contextual (it was present prior to the conduct of the evaluation), I indicated the ensuing evaluator action. Alternatively, when the factor was an evaluator action—I identified the context from which the action emanated.

Let me briefly comment on some of these factor categories. They provide the stimulus for the signature of my prescriptive theory. The first of these is the "evaluator factors" category. This includes such things as the evaluator's personal commitment to use. By this I refer to the way in which the evaluator—in his or her actions and interactions—demonstrates that attaining use is a very high priority. Moreover, it is a priority that the evaluator seeks to infuse into users' thinking. This personal commitment to use and gaining users' attention to this priority is attained by actively teaching stakeholders to be users. This involves,

at minimum, a willingness on the part of evaluators to involve users in the conduct of the evaluation.

Another evaluator factor within this category is of great importance. The evaluator's "credibility" is an important part of the evaluation being subsequently viewed as credible and thus worthy of use. It is important to note that credibility on the part of the evaluator, while partially attained at the outset, is mostly acquired through the actions of the evaluator (we talked about this in Section C).

The grouping called "user factors" includes items such as meaningful participation in the evaluation by users. As noted, I believe that it is the evaluator's task to seek commitment to use on the part of the stakeholders. That is not often the case. The burden falls on a use-committed evaluator through his or her attention to teaching about use to gain that user commitment.

Another area within the evaluation user category is "influence capability." This refers to the potential role that the users might have in making decisions about program change, or their ability to modify attitudes about the program. For a user's commitment to use to have impact, that user must have the capability, directly or indirectly, to influence actions or attitudes.

In the third category, I note that programs exist within their larger organizations and within communities. This I refer to as the "cultural" set of factors. We have discussed these various elements at multiple points throughout this book. As I noted, this consists of the individual context, the organizational/cultural context factors, and the community context. We talked about these in Sections E, G, and H, respectively. Each of these and their constituent parts as well as their rules and expectations exert influence on an evaluation. But discovery of these factors is not enough—attention to these factors by the evaluator as reflected in his or her actions is what is essential. That is why context awareness is repeated (further reflected) in the first set of evaluator factors.

Finally, in regard to the fourth category, the "evaluation factors," I have noted that the context of the evaluation's commitment to use demands that a credible and useful evaluation be produced. To achieve credibility the evaluation process and outcome must demonstrate high technical quality *appropriate to the situation*. Moreover, evaluation information must be considered relevant and essential. For an evaluation to be relevant it must examine questions that are really of concern to users, the measures for determining success must be considered appropriate by users, and the *valuing* process must *be user engaged*. This is important. The relevance of the information provided by the evaluator is further enhanced by an understanding of the existing competing

information—that is, what do users already know and believe? Alas, this is very difficult for the evaluator to determine, but can be potentially acquired by building strong relationships with stakeholders.

Forgive me. I have been too wordy—that's the professor in me coming out—but I thought that background was necessary. Now let's talk about *context-sensitive evaluation* and its signature. I have restated some of these ideas in a more action-oriented format. Let me direct you to Table Y.1, which details the context-sensitive theoretical signature. These are the things that should be done for an evaluation to be considered a context-sensitive evaluation. You will note that all of the elements are stated as evaluator actions. Evaluators are instrumental in shaping users' views and actions. Evaluators direct the engagement of conducting the evaluation. And evaluators must be cognizant of the multidimensions of context in all that they do.

Most of the items in Table Y.1 should be readily apparent to you—we have been discussing them throughout the book. Let me now take

TABLE Y.1. Theoretical Signature of Context-Sensitive Evaluation

1. Evaluator has deep personal commitment to the goal of evaluation use.
2. Evaluator seeks to enhance user's/stakeholder's perception of the value of evaluation use through:
   a. Identifying potential users/stakeholders who have an interest in evaluation use or who are amenable to obtaining that orientation.
   b. Identifying those who have influence capability.
   c. Training users/stakeholders to be users primarily by their engagement and deep participation in the evaluation.
   d. Engaging primary users/stakeholders in reality checking the relevance of questions and measures.
   e. Engaging primary users/stakeholders in *a priori* valuing with reality testing.
3. Evaluator seeks to enhance credibility through demonstrated technical competence.
4. Evaluator seeks to demonstrate professional competence through stakeholder relevance and ongoing commitment to relationship building.
5. Evaluator systematically seeks information about the individual/organizational/community/cultures and considers their relevance for the evaluation.

each of them and provide some *concrete actions* that you might take in order to fulfill each element of the theoretical signature.

1. *Evaluator has deep personal commitment to the goal of evaluation use.* You, as the evaluator, introduce yourself as committed to use. Pay attention to commenting on and emphasizing use in meetings with stakeholders.

2a. *Evaluator seeks to enhance user's/stakeholder's perception of the value of evaluation use through identifying potential users/stakeholders who have an interest in evaluation use or who are amenable to obtaining that orientation.* You should ask questions like: "Why are we doing this evaluation?" "What use might be made of it?" "Would program changes be possible?" "What do you think you might learn about your program?" "What do you hope/think you might learn about evaluation?"

2b. *Evaluator seeks to enhance user's/stakeholder's perception of the value of evaluation use through identifying those who have influence capability.* You should ask: "What role might you have in influencing program change of any type at the conclusion of the evaluation?"

2c. *Evaluator seeks to enhance user's/stakeholder's perception of the value of evaluation use through training users/stakeholders to be users primarily by their engagement and deep participation in the evaluation.* You should engage users/stakeholders in logic modeling with you questioning input/output/outcome relationships. Also, you should be involved in helping users/stakeholders to define questions, determine the relevance of measures, and develop *a priori* valuing schemes. You should actively work with users/stakeholders after findings are presented in considering implications of these findings for change.

2d. *Evaluator seeks to enhance user's/stakeholder's perception of the value of evaluation use through engaging primary users/stakeholders in reality checking the relevance of questions and measures.* You should ask questions like: "Why is this question important?" "What difference would it make?" "Might the answer to that question provide support for program changes or for modified attitudes?" "Would you be satisfied that this measure provides relevant evidence related to the question that you have asked?"

2e. *Evaluator seeks to enhance user's/stakeholder's perception of the value of evaluation use through engaging primary users/stakeholders in* a priori *valuing with reality testing.* You should ask questions like: "If I told you at the end of the evaluation that the level of accomplishment for this particular question was X, would you conclude that was an indication of success?" "What if it was Y?" "What would you consider to be an appropriate level to be deemed successful?" "If the measure is

qualitative, help me to describe the characteristics and general description of 'success.' "

3. *Evaluator seeks to enhance credibility through demonstrated technical competence.* You should display appropriate qualitative, quantitative, and evaluation design skills. These skills are judged based on their *relevance* to the particular context and to the *stakeholder needs*.

4. *Evaluator seeks to demonstrate professional competence through stakeholder relevance and ongoing commitment to relationship building.* You should behave in a professionally appropriate and relevant manner in performing the various evaluation tasks. All work must be performed in a timely manner. You should continue to engage stakeholders in relationship building in order to increase trust and credibility.

5. *Evaluator systematically seeks information about the individual/organizational/community/cultures and considers their relevance for the evaluation.* You should interview all primary users/stakeholders to better understand their values. You should interview various other stakeholders as well. You should systematically seek to gain information about the organizational and management structure, formal and informal. You should gather community data from interviews and informal sources. You should personally go into the community and talk with community members. You should reflect on your own culture in relation to the various individual/program/community cultures. You should seek to conduct evaluation in a manner sensitive to the cultural context.

GAINING ADDITIONAL UNDERSTANDING

Thought Exercise

Now it's time to apply this learning. I hope that you have taken advantage of the opportunities at the end of the sections of this book and either evaluated the RUPAS Program or considered an evaluation of a program of your choosing. Examine again the items in the context-sensitive evaluation theoretical signature and reflect on how you might have modified the evaluation of the RUPAS Program or of the program that you had selected to evaluate.

Reference

Miller, R. L. (2010). Developing standards for empirical examinations of evaluation theory. *American Journal of Evaluation, 31*(3), 390–399.

SECTION

Z

How Can You Embark on a Program to Learn More about Evaluation?

I now offer three suggestions for how you can learn more about evaluation and improve your skills as an evaluator. The first reinforces the notion of further learning described at the end of each section of this book. The second indicates sources of learning that go beyond this book and its suggested readings. And the third anticipates your actions as an evaluator and discusses ways in which you can obtain feedback to improve your skills and your evaluations.

GETTING FEEDBACK ON EVALUATION

One of my favorite lines is from a poem titled "To a Louse," by the Scottish poet Robert Burns: "O wad some Power the giftie gie us to see oursels as ithers see us" (as written in the original). We learn by seeing how others view us or the products of our efforts. There are multiple ways in which the evaluator may benefit from the Robert Burns call to action. To me, perhaps, the most important source of such information are the *stakeholders*. Consider doing a debriefing with stakeholders. Ask them: "Were the questions I asked the right ones?" "If not, why weren't they the right ones?" "How might I have pushed you to get to these preferred questions?" "Did you feel that the measures that we used were appropriate?" "Did you feel that there were problems in data col-

lection?" "What were the problems, and what might we have done differently?" "Were the statistical analyses sufficiently understandable?" "Was the report presented in a fashion that made sense to you?" "How might it have been improved?" "Do you have a better understanding of your program now?" "What are the actions or decisions that you anticipate might be made that have been enhanced through the evaluation findings?" "My aim was to help you to understand the importance of evaluation use. Did I do that?" "What might I do better to help teach you about how to use an evaluation?" "Do you have a better understanding and appreciation of evaluation as a consequence of this evaluation?"

Another informal source of feedback of the more technical aspects of the evaluation than what can be learned from querying stakeholders is by talking to *other evaluators*. Why not consider finding a colleague—preferably a more experienced colleague—and get together for a short discussion about your report? Here, the questions might reflect more technical issues: Were each of the instruments used valid? In what ways might they have not captured the intent of the questions? Were the statistical analyses that were used appropriate? Were the qualitative data sufficiently detailed to justify the conclusions? And so on.

A more formal means of getting feedback on evaluation is through what is referred to as a *metaevaluation*. You will recall that metaevaluation was included within the fifth category of the *Program Evaluation Standards* detailed in Section W. The procedure for conducting a metaevaluation at its simplest is to examine whether the evaluation was performed in accordance with the criteria set forth in the *Standards*. There are some evaluation writers who advocate that each evaluation should have a metaevaluation conducted by an external evaluator—that is, an evaluation (in accord with the *Standards*) of the evaluation. I personally think that this is overkill and simply is not feasible. Typically, we barely have sufficient resources to do an adequate evaluation, let alone establish resources for the conduct of a metaevaluation. But if resources are available, fine. It might be an excellent basis for reassuring stakeholders and providing you with feedback. If not, you might have a colleague at least reflect upon your evaluation in terms of the *Standards*.

TAKING FULL ADVANTAGE OF THIS BOOK

Only a brief discussion of this is necessary. Throughout the book, I have provided opportunities at the end of each section to gain further understanding of evaluation based on the theory underlying my con-

ception of evaluation. If you responded to my suggestions, you would have engaged in further reading in order to increase your knowledge related to the topic.

Two possibilities were presented for practicing evaluation. The RUPAS case study that I presented is a vehicle that you hopefully might have employed for applying acquired skills. If you had decided to engage in a simulated evaluation of a program with which you were familiar, then hopefully the accompanying evaluation and "thought questions" were helpful.

Further, a poem written by a student in one of my classes is presented in Appendix B. You might find it interesting. If you have not been mindful of my invitation to engage in active learning, then let me now present a suggestion. Think about the sections of this book. Go to the outline at the beginning of the book: Which sections are a little hazy? Why not go back and do some further reading?

GAINING EVALUATION EXPERTISE BEYOND THIS BOOK

How do we learn? We read, we practice, we participate, and we listen. In this book, you have engaged in reading—both within the sections as well as in the additional readings. I have not mentioned other evaluation textbooks—my pride of ownership tells me that they are not as readable as *Evaluation Essentials*. But if you wish to learn more and get another perspective, there are certainly interesting books to read.

We also learn by practicing. You have had a simulated opportunity to do some practice based upon the case study—that is, you practiced doing, or at least thinking about doing, an evaluation. You might find it worthwhile to examine other evaluation case situations listed at the end of this section. I reference two books that are worth reading for insights into how experienced evaluators say they conducted (or would conduct) specific evaluations.

We also learn by participating. Evaluation is a profession. There are specific competencies that one anticipates that evaluators would have (see the Further Reading of this section). There are evaluation degree training programs and there are evaluation professional associations throughout the world. Many of these have annual conferences, which are very helpful. One such conference is the annual conference of the American Evaluation Association (AEA). Here you will have the opportunity to listen to many stimulating presentations from leaders in the field. Perhaps you will also consider sharing your own evalua-

tion experiences with others. Details may be found online at the website *eval.org*.

I make it a point each year to attend the AEA conference. Since you and I are not strangers—indeed, how could we be? We have had a long conversation—feel free to come up and say hello.

GAINING ADDITIONAL UNDERSTANDING

Further Reading

Robinson, S. B. (2011). Inside, outside, upside down: Challenges and opportunities that frame the future of a novice evaluator. *New Directions for Evaluation, 131,* 65–70.

You will find one evaluator's thoughtful reflections on potential growth opportunities in evaluation in this article.

Stevahn, L., King, J. A., Ghere, G., & Minnema, J. (2005). Establishing essential competencies for program evaluators. *American Journal of Evaluation, 26*(1), 43–59.

If taxonomies are helpful to your learning, then you will find a comprehensive one on essential evaluator competencies in this article.

Quick Reads

1. Lisa Dillman on Taking Control of Your Evaluation Career
   http://tinyurl.com/jpszlj5

2. Çigdem Meek, Bashar Ahmed, and Marissa Molina on Lessons Learned as Novice Evaluators: Eval2014
   http://tinyurl.com/zzwxbny

3. Laura Sundstrom and Megan Elyse Williams on Intentionally Developing Skills in New Evaluators: The Tier Approach
   http://tinyurl.com/hh4zo2l

Studying Cases

I invite you here to learn about evaluation by studying cases. Here are two excellent sources that you might examine.

Alkin, M. C., & Christie, C. A. (Eds.). (2005). *Theorists' models in action* (New Directions in Evaluation No. 106). San Francisco: Jossey-Bass.

In this monograph, four different evaluators describe how they would do an evaluation of a public–private enterprise—the Bunche–Da Vinci Partnership.

Fitzpatrick, J., Christie, C. A., & Mark, M. M. (2009). *Evaluation in action: Interviews with expert evaluators.* Thousand Oaks, CA: SAGE.

In this excellent volume, Fitzpatrick and her colleagues provide a brief description of each of 12 different evaluations and then interview the evaluators to answer questions about why they did what they did.

Real-World Experience

Now, perhaps, you are ready to go out and assist on an evaluation. Or, possibly, you might have an opportunity to be an active observer of an ongoing evaluation. Try it!

APPENDIX A

Factors Affecting Evaluation Use

Evaluator factors		User factors	
Context	**Evaluator actions**	**Context**	**Evaluator actions**
1. Commitment to use	• Willingness to involve users • Relationship building	1. Evaluator commitment to use	• Develop stakeholder commitment to use • Engage stakeholder in meaningful participation
2. Evaluator technical competence	• Evaluator further develops these competencies.	2. Initial attitude toward evaluators	• Evaluator works on developing in users a positive attitude toward evaluation.
3. Initial evaluator credibility	• Evaluator works toward building greater credibility.	3. User initial influence capability	• Evaluator identifies those who have influence capability.
4. Individual/ organizational/ personal context	• Evaluator works toward developing further awareness of and sensitivity to: a. Individual cultural contexts b. Organizational and cultural context factors c. Self in relationship to the various contexts	4. Initial evaluator–user relationship	• Evaluator engages in ongoing communication and relationship building.

Organizational/cultural context factors		Evaluation factors	
Context	**Evaluator actions**	**Context**	**Evaluator actions**
1. Preexisting evaluation political bounds a. Written evaluation requirements b. Other contractual obligations c. Organizational evaluation policy d. Fiscal constraints	• Evaluator seeks to understand these bounds.	Evaluator commitment to use	• Evaluator engages in an evaluation that is credible and useful. • Technical quality (appropriate to the situation) • Relevance (meets information needs) • Understands relationships to existing and competing information • High communication quality and appropriateness • Timeliness • Structured options for possible use understanding presented
2. Organizational and programmatic arrangements a. Development stage of the program b. Interest in evaluation from higher organizational level	• Evaluator accommodates the evaluation to the stage of the program. • Evaluator seeks to understand these interests and use them to enhance the evaluation.		
3. Interorganizational elements a. Relationship to the broader program b. Relationship to related programs in the organization	• Evaluator seeks to understand and be sure that these contexts are reflected in the evaluation.		
4. Relationship to and impact on the relevant stakeholder and community/cultural context	• Evaluator seeks to understand and be sure that these contexts are reflected in the evaluation.		

APPENDIX B

An Evaluation Lesson

by "Unknown Student"

Get your control group
Make an experimental design
Randomly choose all your subjects
Allow plenty of time

Pretest and posttest
Use accuracy and care
Any unanticipated outcomes
You must declare

Train all your staff
And visit the site
Now the perfect evaluation
Will be done just right

The data are collected
Analyzed with expertise
The report carefully written
Is sure to please

Now you sit back and await
After your tiring escapade
The news from the boss
"Program changes will be made."

You have not long to wait
For the phone quickly rings
Your presence is asked for
To clarify a few things

This poem was written by a student in one of Marv's classes many years ago and given to him as a gift. Unfortunately, the author's name was not included and all attempts to search records have been unsuccessful. We would be very pleased to have the author contact us to get reacquainted and to include her name in subsequent printings.

"It's interesting," you're told
"But not what we need
The goals are different
Than we agreed."

"But, but," you sputter
"It should be done this way
You guys are not in tune
With evaluation today!"

"Sorry sir," you're told
"It is you and not us
We don't need all this data
of T-tests and stuff

The things you found out
Will not help our lot
We need to find ways
To improve what we got.

We can't change our teachers
We can't change our hours
We can't change our SES
We don't have that power.

We called you to help
Evaluate what we're doing
Not judge all our faults
Rather suggest ways of improving."

A comment is made
As you leave the door
That you return to school
And learn some more

Evaluation, they say
Is not research design.
Randomization and control groups
Needn't be used all the time

They said you begin
By first finding out
Who makes the decisions
And what they are about

Next, ask relevant questions
In order to know
Many more things
Than just "go or no"

Find out why they wish
An evaluation done
And who are the people
They think must be won?

You must address, too
In your report
A number of groups—
It's quite a sport!

You should involve yourself
And take a good look
At the actual program
Not the one in the book

Study as a whole program
As well as the parts
Observe, interview, question
It's all part of the art

Don't go blowing whistles
On people or goals
You're not a policeman
For that's not your role

Keep in close contact
With those who asked for advice
And make your report
Helpful, clear, and concise

Don't expect miracles
Because you did a good job
You provided more data
To add to the blob

Utilization will occur
In small little bits
And your evaluation report
Helps decisions better fit

This is just a small part
Of the evaluator's role
Take Dr. Alkin's #460
To learn of your goal

Perhaps you won't be able
The perfect report to complete
But you'll know for sure
How to tackle the feat

You'll learn it's an art
And not just a skill
The evaluator's job
Is more than research drill

"Come back," you're told
"When these things you learn
And if you're serious about evaluation
We'll give you another turn."

APPENDIX C

Use Factors: Relationship to Research Compilations

Factors affecting evaluation use	References*				
	Leviton & Hughes (1981)	Alkin (1985)	Cousins & Leithwood (1986)	Shulha & Cousins (1997)	Johnson et al. (2009)
EvaluaTOR factors					
1. Commitment to use		✓	✓	✓	✓
2. Willingness to involve users		✓	✓	✓	✓
3. Technical competence	✓	✓	✓		✓
4. Credibility	✓	✓	✓	✓	✓
5. Context awareness		✓		✓	
a. Culture					✓
b. Relationship building		✓		✓	
c. Political sensitivity	✓	✓	✓	✓	✓
d. Awareness of self in relationship to context				✓	

(continued)

APPENDIX C *(continued)*

Factors affecting evaluation use	References*				
	Leviton & Hughes (1981)	**Alkin (1985)**	**Cousins & Leithwood (1986)**	**Shulha & Cousins (1997)**	**Johnson et al. (2009)**
User factors					
1. Commitment to use	✓	✓	✓	✓	✓
2. Meaningful participation	✓	✓		✓	✓
3. Attitude toward evaluation	✓	✓	✓	✓	✓
4. Previous experience in evaluation		✓	✓		
5. Previous evaluation success	✓				
6. Influence capability		✓			✓
EvaluaTION factors					
1. Evaluation credibility	✓	✓	✓	✓	✓
a. Technical quality (appropriate to the situation)	✓	✓	✓	✓	✓
b. Relevance (meets information needs)	✓	✓	✓	✓	✓
c. Relationships to existing/competing information	✓		✓	✓	✓
d. Communication quality and appropriateness	✓	✓	✓	✓	✓
e. Timeliness	✓	✓	✓		✓
f. Structured options for possible use understanding		✓			
2. Ongoing communication and relationship building	✓	✓		✓	

(continued)

APPENDIX C *(continued)*

Factors affecting evaluation use	References*				
	Leviton & Hughes (1981)	Alkin (1985)	Cousins & Leithwood (1986)	Shulha & Cousins (1997)	Johnson et al. (2009)
Organizational/social context factors					
1. Preexisting evaluation bounds		✓			
a. Written evaluation requirements		✓			
b. Other contractual obligations		✓			✓
c. Organizational evaluation policy	✓				
d. Fiscal constraints		✓	✓		✓
2. Organizational and programmatic arrangements		✓	✓	✓	
a. Development state of the program		✓			
3. Interorganizational elements	✓	✓	✓		
a. Relationship to the broader program	✓	✓			✓
b. Relationship to related programs in the organization		✓		✓	
c. Interest in evaluation from higher organizational level	✓	✓		✓	
4. Relationship to and impact on the relevant stakeholder community/populace	✓	✓		✓	✓

*References: Alkin, M. C. (1985). *A guide for evaluation decision makers.* Beverly Hills, CA: Sage. Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56*(3), 331–364. Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation, 30*(3), 377–410. Leviton, L. C., & Hughes, E. F. (1981). Research on the utilization of evaluations: A review and synthesis. *Evaluation Review, 5*(4), 525–548. Shulha, L. M., & Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1986. *Evaluation Practice, 18*(3), 195–208.

Index

Page numbers followed by *f* indicate figures, *t* indicate tables

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z

About the Authors

Marvin C. Alkin, EdD, is Professor Emeritus in the Social Research Methodology Division of the Graduate School of Education and Information Studies at the University of California, Los Angeles (UCLA). He has been a member of the UCLA faculty since 1964, and at various times has served as Chair of the Education Department and Associate Dean of the School. Dr. Alkin is a founder and former Director of the Center for the Study of Evaluation, which was established in 1966 by the U.S. government and which continues to be an integral part of the UCLA Graduate School of Education and Information Studies. A leading authority in the field of evaluation, he is best known for his work on the use of evaluation information in decision making and on comparative evaluation theory. Dr. Alkin is a recipient of the Paul F. Lazarsfeld Evaluation Theory Award and the Research on Evaluation Award from the American Evaluation Association. He has published more than 150 journal articles, book chapters, books, and technical reports; has consulted to numerous national governments; and has directed over 100 program evaluations in the United States and internationally.

Anne T. Vo, PhD, is Assistant Professor of Clinical Medical Education at the Keck School of Medicine of the University of Southern California (USC). Dr. Vo's substantive interests as an evaluation scholar–practitioner lie at the intersection of comparative evaluation theory, evaluation capacity building, and organizational development. Her research, publications, and evaluation practice contribute to the field's

understanding of how evaluation can be practiced better, where and how social science theory and evaluation science dovetail into each other, and how this knowledge can be leveraged to drive change. Dr. Vo established the Keck Evaluation, Institutional Reporting/Research, and Assessment Office at USC, which aims to support the transformation of medical training through research, service, and education, and currently serves as Associate Director. She holds leadership positions in the American Educational Research Association and the American Evaluation Association and serves as Editor of the section on Teaching and Learning Evaluation of the *American Journal of Evaluation*.