4th Edition

Program Evaluation

ALTERNATIVE APPROACHES AND PRACTICAL GUIDELINES

Jody L. Fitzpatrick • James R. Sanders • Blaine R. Worthen

Program Evaluation

Alternative Approaches and Practical Guidelines

FOURTH EDITION

Jody L. Fitzpatrick
University of Colorado Denver

James R. Sanders
Western Michigan University

Blaine R. Worthen
Utah State University

PEARSON

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Vice President and Editor in Chief: Jeffery W. Johnston
Senior Acquisitions Editor: Meredith D. Fossel
Editorial Assistant: Nancy Holstein
Vice President, Director of Marketing: Margaret Waples
Senior Marketing Manager: Christopher D. Barry
Senior Managing Editor: Pamela D. Bennett
Senior Project Manager: Linda Hillis Bayma
Senior Operations Supervisor: Matthew Ottenweller
Senior Art Director: Diane Lorenzo
Cover Designer: Jeff Vanik
Cover Image: istock
Full-Service Project Management: Ashley Schneider, S4Carlisle Publishing Services
Composition: S4Carlisle Publishing Services
Printer/Binder: Courier/Westford
Cover Printer: Lehigh-Phoenix Color/Hagerstown
Text Font: Meridien

Credits and acknowledgments borrowed from other sources and reproduced, with permission, in this textbook appear on appropriate page within text.

Every effort has been made to provide accurate and current Internet information in this book. However, the Internet and information posted on it are constantly changing, so it is inevitable that some of the Internet addresses listed in this textbook will change.

Copyright © 2011, 2004, 1997 Pearson Education, Inc., Upper Saddle River, New Jersey 07458. All rights reserved. Manufactured in the United States of America. This publication is protected by Copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. To obtain permission(s) to use material from this work, please submit a written request to Pearson Education, Inc., Permissions Department, 501 Boylston Street, Suite 900, Boston, MA 02116, fax: (617) 671-2290, email: permissionsus@pearson.com.

Library of Congress Cataloging-in-Publication Data

Fitzpatrick, Jody L.
Program evaluation: alternative approaches and practical guidelines / Jody L. Fitzpatrick, James R. Sanders, Blaine R. Worthen.
p. cm.
ISBN 978-0-205-57935-8
1. Educational evaluation—United States. 2. Evaluation research (Social action programs)—United States. 3. Evaluation—Study and teaching—United States. I. Sanders, James R. II. Worthen, Blaine R. III. Worthen, Blaine R. Program evaluation. IV. Title.
LB2822.75.W67 2011
379.1'54—dc22

2010025390

10 9 8 7 6 5 4 3 2

PEARSON

www.pearsonhighered.com

ISBN 10: 0-205-57935-3
ISBN 13: 978-0-205-57935-8

About the Authors

Jody Fitzpatrick has been a faculty member in public administration at the University of Colorado Denver since 1985. She teaches courses in research methods and evaluation, conducts evaluations in many schools and human service settings, and writes extensively about the successful practice of evaluation. She has served on the Board of the American Evaluation Association and on the editorial boards of the *American Journal of Evaluation* and *New Directions for Evaluation.* She has also served as Chair of the Teaching of Evaluation Topical Interest Group at the American Evaluation Association and has won a university-wide teaching award at her university. In one of her recent publications, *Evaluation in Action: Interviews with Expert Evaluators*, she uses interviews with expert evaluators on one evaluation to talk about the decisions that evaluators face as they plan and conduct evaluations and the factors that influence their choices. She is currently evaluating the changing roles of counselors in middle schools and high schools and a program to help immigrant middle-school girls to achieve and stay in school. Her international work includes research on evaluation in Spain and Europe and, recently, she has spoken on evaluation issues to policymakers and evaluators in France, Spain, Denmark, Mexico, and Chile.

James Sanders is Professor Emeritus of Educational Studies and the Evaluation Center at Western Michigan University where he has taught, published, consulted, and conducted evaluations since 1975. A graduate of Bucknell University and the University of Colorado, he has served on the Board and as President of the American Evaluation Association (AEA) and has served as Chair of the Steering Committee that created the Evaluation Network, a predecessor to AEA. His publications include books on school, student, and program evaluation. He has worked extensively with schools, foundations, and government and nonprofit agencies to develop their evaluation practices. As Chair of the Joint Committee on Standards for Educational Evaluation, he led the development of the second edition of *The Program Evaluation Standards.* He was also involved in developing the concepts of *applied performance testing* for student assessments, *cluster evaluation* for program evaluations by foundations and government agencies, and *mainstreaming evaluation* for organizational development. His international work in evaluation has been concentrated in Canada, Europe, and Latin America. He received distinguished service awards from Western Michigan University, where he helped to establish a PhD program in evaluation, and from the Michigan Association for Evaluation.

Blaine Worthen is Psychology Professor Emeritus at Utah State University, where he founded and directed the Evaluation Methodology PhD program and the Western Institute for Research and Evaluation, conducting more than 350 evaluations for local and national clients in the United States and Canada. He received his PhD from The Ohio State University. He is a former editor of *Evaluation Practice* and founding editor of the *American Journal of Evaluation.* He served on the American Evaluation Association Board of Directors and received AEA's Myrdal Award for Outstanding Evaluation Practitioner and AERA's Best Evaluation Study Award. He has taught university evaluation courses (1969–1999), managed federally mandated evaluations in 17 states (1973–1978), advised numerous government and private agencies, and given more than 150 keynote addresses and evaluation workshops in the United States, England, Australia, Israel, Greece, Ecuador, and other countries. He has written extensively in evaluation, measurement, and assessment and is the author of 135 articles and six books. His *Phi Delta Kappan* article, "Critical Issues That Will Determine the Future of Alternative Assessment," was distributed to 500 distinguished invitees at the White House's Goals 2000 Conference. He is recognized as a national and international leader in the field.

Preface

The twenty-first century is an exciting time for evaluation. The field is growing. People—schools, organizations, policymakers, the public at large—are interested in learning more about how programs work: how they succeed and how they fail. Given the tumult experienced in the first decade of this century, many people are interested in accountability from corporations, government, schools, and nonprofit organizations. The fourth edition of our best-selling textbook is designed to help readers consider how evaluation can achieve these purposes. As in previous editions, our book is one of the few to introduce readers to both the different approaches to evaluation and practical methods for conducting it.

New to This Edition

The fourth edition includes many changes:

- A new chapter on the role of politics in evaluation and ethical considerations.
- A new and reorganized Part Two that presents and discusses the most current approaches and theories of evaluation.
- An increased focus on mixed methods in design, data collection, and analysis.
- Links to interviews with evaluators who conducted an evaluation that illustrates the concepts reviewed in that chapter, as they discuss the choices and challenges they faced.
- A discussion of how today's focus on performance measurement, outcomes, impacts, and standards have influenced evaluation.
- New sections on organizational learning, evaluation capacity building, mainstreaming evaluation, and cultural competence—trends in evaluation and organizations.

Evaluation, today, is changing in a variety of ways. Policymakers, managers, citizens, and consumers want better tracking of activities and outcomes. More importantly, many want a better *understanding* of social problems and the programs and policies being undertaken to reduce these problems. Evaluation in many forms, including performance measurement and outcome or impact assessments, is expanding around the globe. People who work in organizations are also interested in evaluation as a way to enhance organizational learning. They want to know how well they're doing, how to tackle the tough problems their organizations address, and how to improve their performance and better serve their clients and their

community. Many different methods are being developed and used: mixed methods for design and data collection, increased involvement of new and different stakeholders in the evaluation process, expanded consideration of the potential uses and impacts of evaluation, and more effective and diverse ways to communicate findings. As evaluation expands around the world, the experiences of adapting evaluation to different settings and different cultures are enriching the field.

In this new edition, we hope to convey to you the dynamism and creativity involved in conducting evaluation. Each of us has many years of experience in conducting evaluations in a variety of settings, including schools, public welfare agencies, mental health organizations, environmental programs, nonprofit organizations, and corporations. We also have years of experience teaching students how to use evaluation in their own organizations or communities. Our goal is, and always has been, to present information that readers can use either to conduct or to be a participant in evaluations that make a difference to their workplace, their clients, and their community. Let us tell you a bit more about how we hope to do that in this new edition.

Organization of This Text

The book is organized in four parts. Part One introduces the reader to key concepts in evaluation; its history and current trends; and ethical, political, and interpersonal factors that permeate and transcend all phases of evaluation. Evaluation differs from research in that it is occurring in the real world with the goal of being used by non-researchers to improve decisions, governance, and society. As a result, evaluators develop relationships with their users and stakeholders and work in a political environment in which evaluation results compete with other demands on decision makers. Evaluators must know how to work in such environments to get their results used. In addition, ethical challenges often present themselves. We find the ways in which evaluation differs from research to be both challenging and interesting. It is why we chose evaluation as our life's work. In Part One, we introduce you to these differences and to the ways evaluators work in this public, political context.

In Part Two, we present several different approaches, often called models or theories, to evaluation. (Determining whether objectives or outcomes have been achieved isn't the only way to approach evaluation!) Approaches influence how evaluators determine what to study and how they involve others in what they study. We have expanded our discussions of theory-based, decision-oriented, and participatory approaches. In doing so, we describe new ways in which evaluators use logic models and program theories to understand the workings of a program. Participatory and transformative approaches to empowering stakeholders and creating different ways of learning are described and contrasted. Evaluators must know methodology, but they also must know about different approaches to evaluation to consciously and intelligently choose the approach or mix of approaches that is most appropriate for the program, clients, and stakeholders and context of their evaluation.

In Parts Three and Four, the core of the book, we describe how to plan and carry out an evaluation study. Part Three is concerned with the planning stage: learning about the program, conversing with stakeholders to learn purposes and consider future uses of the study, and identifying and finalizing evaluation questions to guide the study. Part Three teaches the reader how to develop an evaluation plan and a management plan, including timelines and budgets for conducting the study. In Part Four, we discuss the methodological choices and decisions evaluators make: selecting and developing designs; sampling, data collection, and analysis strategies; interpreting results; and communicating results to others. The chapters in each of these sections are sequential, representing the order in which decisions are made or actions are taken in the evaluation study. We make use of extensive graphics, lists, and examples to illustrate practice to the reader.

This Revision

Each chapter has been revised by considering the most current books, articles, and reports. Many new references and contemporary examples have been added. Thus, readers are introduced to current controversies about randomized control groups and appropriate designs for outcome evaluations, current discussions of political influences on evaluation policies and practices, research on participative approaches, discussions of cultural competency and capacity building in organizations, and new models of evaluation use and views on interpreting and disseminating results.

We are unabashedly eclectic in our approach to evaluation. We use many different approaches and methods—whatever is appropriate for the setting—and encourage you to do the same. We don't advocate one approach, but instruct you in many. You will learn about different approaches or theories in Part Two and different methods of collecting data in Parts Three and Four.

To facilitate learning, we have continued with much the same pedagogical structure that we have used in past editions. Each chapter presents information on current and foundational issues in a practical, accessible manner. Tables and figures are used frequently to summarize or illustrate key points. Each chapter begins with Orienting Questions to introduce the reader to some of the issues that will be covered in the chapter and concludes with a list of the Major Concepts and Theories reviewed in the chapter, Discussion Questions, Application Exercises, and a list of Suggested Readings on the topics discussed.

Rather than using the case study method from previous editions, we thought it was time to introduce readers to some real evaluations. Fortunately, while Blaine Worthen was editor of *American Journal of Evaluation,* Jody Fitzpatrick wrote a column in which she interviewed evaluators about a single evaluation they had conducted. These interviews are now widely used in teaching about evaluation. We have incorporated them into this new edition by recommending the ones that illustrate the themes introduced in each chapter. Readers and instructors can choose either to purchase the book, *Evaluation in Action* (Fitzpatrick, Christie, & Mark, 2009), as a case companion to this text or to access many of the interviews

through their original publication in the *American Journal of Evaluation*. At the end of each chapter, we describe one to three relevant interviews, citing the chapter in the book and the original source in the journal.

We hope this book will inspire you to think in a new way about issues—in a questioning, exploring, evaluative way—and about programs, policy, and organizational change. For those readers who are already evaluators, this book will provide you with new perspectives and tools for your practice. For those who are new to evaluation, this book will make you a more informed consumer of or participant in evaluation studies or, perhaps, guide you to undertake your own evaluation.

Acknowledgments

We would like to thank our colleagues in evaluation for continuing to make this such an exciting and dynamic field! Our work in each revision of our text has reminded us of the progress being made in evaluation and the wonderful insights of our colleagues about evaluation theory and practice. We would also like to thank Sophia Le, our research assistant, who has worked tirelessly, creatively, and diligently to bring this manuscript to fruition. We all are grateful to our families for the interest and pride they have shown in our work and the patience and love they have demonstrated as we have taken the time to devote to it.

Contents

PART ONE • *Introduction to Evaluation* **1**

1 *Evaluation's Basic Purpose, Uses, and Conceptual Distinctions* **3**

Informal versus Formal Evaluation **5**
A Brief Definition of Evaluation and Other Key Terms **6**
Differences in Evaluation and Research **9**
The Purposes of Evaluation **13**
Roles and Activities of Professional Evaluators **16**
Uses and Objects of Evaluation **18**
Some Basic Types of Evaluation **20**
Evaluation's Importance—and Its Limitations **32**

2 *Origins and Current Trends in Modern Program Evaluation* **38**

The History and Influence of Evaluation in Society **38**
1990–The Present: History and Current Trends **49**

3 *Political, Interpersonal, and Ethical Issues in Evaluation* **64**

Evaluation and Its Political Context **65**
Maintaining Ethical Standards: Considerations, Issues, and Responsibilities for Evaluators **78**

PART TWO • ***Alternative Approaches to Program Evaluation*** **109**

4 ***Alternative Views of Evaluation*** **111**

Diverse Conceptions of Program Evaluation **113**

Origins of Alternative Views of Evaluation **114**

Classifications of Evaluation Theories or Approaches **120**

5 ***First Approaches: Expertise and Consumer-Oriented Approaches*** **126**

The Expertise-Oriented Approach **127**

The Consumer-Oriented Evaluation Approach **143**

6 ***Program-Oriented Evaluation Approaches*** **153**

The Objectives-Oriented Evaluation Approach **154**

Logic Models and Theory-Based Evaluation Approaches **159**

How Program-Oriented Evaluation Approaches Have Been Used **164**

Strengths and Limitations of Program-Oriented Evaluation Approaches **166**

Goal-Free Evaluation **168**

7 ***Decision-Oriented Evaluation Approaches*** **172**

Developers of Decision-Oriented Evaluation Approaches and Their Contributions **173**

The Decision-Oriented Approaches **173**

How the Decision-Oriented Evaluation Approaches Have Been Used **184**

Strengths and Limitations of Decision-Oriented Evaluation Approaches **184**

8 *Participant-Oriented Evaluation Approaches* **189**

Evolution of Participatory Approaches **190**

Developers of Participant-Oriented Evaluation Approaches and Their Contributions **191**

Participatory Evaluation Today: Two Streams and Many Approaches **199**

Some Specific Contemporary Approaches **205**

How Participant-Oriented Evaluation Approaches Have Been Used **220**

Strengths and Limitations of Participant-Oriented Evaluation Approaches **223**

9 *Other Current Considerations: Cultural Competence and Capacity Building* **231**

The Role of Culture and Context in Evaluation Practice and Developing Cultural Competence **232**

Evaluation's Roles in Organizations: Evaluation Capacity Building and Mainstreaming Evaluation **235**

10 *A Comparative Analysis of Approaches* **243**

A Summary and Comparative Analysis of Evaluation Approaches **243**

Cautions About the Alternative Evaluation Approaches **244**

Contributions of the Alternative Evaluation Approaches **248**

Comparative Analysis of Characteristics of Alternative Evaluation Approaches **249**

Eclectic Uses of the Alternative Evaluation Approaches **251**

PART THREE • *Practical Guidelines for Planning Evaluations* **257**

11 *Clarifying the Evaluation Request and Responsibilities* **259**

Understanding the Reasons for Initiating the Evaluation **260**

Conditions Under Which Evaluation Studies Are Inappropriate **265**

Determining When an Evaluation Is Appropriate: Evaluability Assessment **268**

Using an Internal or External Evaluator **271**

Hiring an Evaluator **277**

How Different Evaluation Approaches Clarify the Evaluation Request and Responsibilities **281**

12 *Setting Boundaries and Analyzing the Evaluation Context* **286**

Identifying Stakeholders and Intended Audiences for an Evaluation **287**

Describing What Is to Be Evaluated: Setting the Boundaries **290**

Analyzing the Resources and Capabilities That Can Be Committed to the Evaluation **304**

Analyzing the Political Context for the Evaluation **307**

Variations Caused by the Evaluation Approach Used **309**

Determining Whether to Proceed with the Evaluation **310**

13 *Identifying and Selecting the Evaluation Questions and Criteria* **314**

Identifying Useful Sources for Evaluation Questions: The Divergent Phase **315**

Selecting the Questions, Criteria, and Issues to Be Addressed: The Convergent Phase **328**

Specifying the Evaluation Criteria and Standards **332**

Remaining Flexible during the Evaluation: Allowing New Questions, Criteria, and Standards to Emerge **336**

14 *Planning How to Conduct the Evaluation* **340**

Developing the Evaluation Plan **342**

Specifying How the Evaluation Will Be Conducted: The Management Plan **358**

Establishing Evaluation Agreements and Contracts 367
Planning and Conducting the Metaevaluation 368

PART FOUR • ***Practical Guidelines for Conducting and Using Evaluations*** **379**

15 ***Collecting Evaluative Information: Design, Sampling, and Cost Choices*** **381**

Using Mixed Methods 383
Designs for Collecting Descriptive and Causal Information 387
Sampling 407
Cost Analysis 411

16 ***Collecting Evaluative Information: Data Sources and Methods, Analysis, and Interpretation*** **418**

Common Sources and Methods for Collecting Information 419
Planning and Organizing the Collection of Information 443
Analysis of Data and Interpretation of Findings 444

17 ***Reporting Evaluation Results: Maximizing Use and Understanding*** **453**

Purposes of Evaluation Reporting and Reports 454
Different Ways of Reporting 455
Important Factors in Planning Evaluation Reporting 456
Key Components of a Written Report 469
Suggestions for Effective Oral Reporting 476
A Checklist for Good Evaluation Reports 479
How Evaluation Information Is Used 479

18 *The Future of Evaluation* **490**

The Future of Evaluation **490**
Predictions Concerning the Profession of Evaluation **491**
Predictions Concerning the Practice of Evaluation **493**
A Vision for Evaluation **496**
Conclusion **497**

Appendix A The Program Evaluation Standards and Guiding Principles for Evaluators 499

References 505

Author Index 526

Subject Index 530

Part I

Introduction to Evaluation

This initial section of our text provides the background necessary for the beginning student to understand the chapters that follow. In it, we attempt to accomplish three things: to explore the concept of evaluation and its various meanings, to review the history of program evaluation and its development as a discipline, and to introduce the reader to some of the factors that influence the practice of evaluation. We also acquaint the reader with some of the current controversies and trends in the field.

In Chapter 1, we discuss the basic purposes of evaluation and the varying roles evaluators play. We define evaluation specifically, and we introduce the reader to several different concepts and distinctions that are important to evaluation. In Chapter 2, we summarize the origins of today's evaluation tenets and practices and the historical evolution of evaluation as a growing force in improving our society's public, nonprofit, and corporate programs. In Chapter 3, we discuss the political, ethical, and interpersonal factors that underlie any evaluation and emphasize its distinction from research.

Our intent in Part One is to provide the reader with information essential to understanding not only the content of the sections that follow but also the wealth of material that exists in the literature on program evaluation. Although the content in the remainder of this book is intended to apply primarily to the evaluation of programs, most of it also applies to the evaluation of policies, products, and processes used in those areas and, indeed, to any object of an evaluation. In Part Two we will introduce you to different approaches to evaluation to enlarge your understanding of the diversity of choices that evaluators and stakeholders make in undertaking evaluation.

This page intentionally left blank

1

Evaluation's Basic Purpose, Uses, and Conceptual Distinctions

Orienting Questions

1. What is evaluation? Why is it important?
2. What is the difference between formal and informal evaluation?
3. What are some purposes of evaluation? What roles can the evaluator play?
4. What are the major differences between formative and summative evaluations?
5. What questions might an evaluator address in a needs assessment, a process evaluation, and an outcome evaluation?
6. What are the advantages and disadvantages of an internal evaluator? An external evaluator?

The challenges confronting our society in the twenty-first century are enormous. Few of them are really new. In the United States and many other countries, the public and nonprofit sectors are grappling with complex issues: educating children for the new century; reducing functional illiteracy; strengthening families; training people to enter or return to the workforce; training employees who currently work in an organization; combating disease and mental illness; fighting discrimination; and reducing crime, drug abuse, and child and spouse abuse. More recently, pursuing and balancing environmental and economic goals and working to ensure peace and economic growth in developing countries have become prominent concerns. As this book is written, the United States and many countries around

the world are facing challenging economic problems that touch every aspect of society. The policies and programs created to address these problems will require evaluation to determine which solutions to pursue and which programs and policies are working and which are not. Each new decade seems to add to the list of challenges, as society and the problems it confronts become increasingly complex.

As society's concern over these pervasive and perplexing problems has intensified, so have its efforts to resolve them. Collectively, local, regional, national, and international agencies have initiated many programs aimed at eliminating these problems or their underlying causes. In some cases, specific programs judged to have been ineffective have been "mothballed" or sunk outright, often to be replaced by a new program designed to attack the problem in a different—and, hopefully, more effective—manner.

In more recent years, scarce resources and budget deficits have posed still more challenges as administrators and program managers have had to struggle to keep their most promising programs afloat. Increasingly, policymakers and managers have been faced with tough choices, being forced to cancel some programs or program components to provide sufficient funds to start new programs, to continue others, or simply to keep within current budgetary limits.

To make such choices intelligently, policy makers need good information about the relative effectiveness of programs. Which programs are working well? Which are failing? What are the programs' relative costs and benefits? Similarly, each program manager needs to know how well different parts of programs are working. What can be done to improve those parts of the program that are not working as well as they should? Have all aspects of the program been thought through carefully at the planning stage, or is more planning needed? What is the theory or logic model for the program's effectiveness? What adaptations would make the program more effective?

Answering such questions is the major task of program evaluation. The major task of this book is to introduce you to evaluation and the vital role it plays in virtually every sector of modern society. However, before we can hope to convince you that good evaluation is an essential part of good programs, we must help you understand at least the basic concepts in each of the following areas:

- How we—and others—define evaluation
- How formal and informal evaluation differ
- The basic purposes—and various uses—of formal evaluation
- The distinction between basic types of evaluation
- The distinction between internal and external evaluators
- Evaluation's importance and its limitations

Covering all of those areas thoroughly could fill a whole book, not just one chapter of an introductory text. In this chapter, we provide only brief coverage of each of these topics to orient you to concepts and distinctions necessary to understand the content of later chapters.

Informal versus Formal Evaluation

Evaluation is not a new concept. In fact, people have been evaluating, or examining and judging things, since the beginning of human history. Neanderthals practiced it when determining which types of saplings made the best spears, as did Persian patriarchs in selecting the most suitable suitors for their daughters, and English yeomen who abandoned their own crossbows in favor of the Welsh longbow. They had observed that the longbow could send an arrow through the stoutest armor and was capable of launching three arrows while the crossbow sent only one. Although no formal evaluation reports on bow comparisons have been unearthed in English archives, it is clear that the English evaluated the longbow's value for their purposes, deciding that its use would strengthen them in their struggles with the French. So the English armies relinquished their crossbows, perfected and improved on the Welsh longbow, and proved invincible during most of the Hundred Years' War.

By contrast, French archers experimented briefly with the longbow, then went back to the crossbow—and continued to lose battles. Such are the perils of poor evaluation! Unfortunately, the faulty judgment that led the French to persist in using an inferior weapon represents an informal evaluation pattern that has been repeated too often throughout history.

As human beings, we evaluate every day. Practitioners, managers, and policymakers make judgments about students, clients, personnel, programs, and policies. These judgments lead to choices and decisions. They are a natural part of life. A school principal observes a teacher working in the classroom and forms some judgments about that teacher's effectiveness. A program officer of a foundation visits a substance abuse program and forms a judgment about the program's quality and effectiveness. A policymaker hears a speech about a new method for delivering health care to uninsured children and draws some conclusions about whether that method would work in his state. Such judgments are made every day in our work. These judgments, however, are based on informal, or unsystematic, evaluations.

Informal evaluations can result in faulty or wise judgments. But, they are characterized by an absence of breadth and depth because they lack systematic procedures and formally collected evidence. As humans, we are limited in making judgments both by the lack of opportunity to observe many different settings, clients, or students and by our own past experience, which both informs and biases our judgments. Informal evaluation does not occur in a vacuum. Experience, instinct, generalization, and reasoning can all influence the outcome of informal evaluations, and any or all of these may be the basis for sound, or faulty, judgments. Did we see the teacher on a good day or a bad one? How did our past experience with similar students, course content, and methods influence our judgment? When we conduct informal evaluations, we are less cognizant of these limitations. However, when formal evaluations are not possible, informal evaluation carried out by knowledgeable, experienced, and fair people can be very useful indeed. It would be unrealistic to think any individual, group, or organization could formally evaluate everything it does. Often informal evaluation is the only

practical approach. (In choosing an entrée from a dinner menu, only the most compulsive individual would conduct exit interviews with restaurant patrons to gather data to guide that choice.)

Informal and formal evaluation, however, form a continuum. Schwandt (2001a) acknowledges the importance and value of everyday judgments and argues that evaluation is not simply about methods and rules. He sees the evaluator as helping practitioners to "cultivate critical intelligence." Evaluation, he notes, forms a middle ground "between overreliance on and over-application of method, general principles, and rules to making sense of ordinary life on one hand, and advocating trust in personal inspiration and sheer intuition on the other" (p. 86). Mark, Henry, and Julnes (2000) echo this concept when they describe evaluation as a form of assisted sense-making. Evaluation, they observe, "has been developed to assist and extend natural human abilities to observe, understand, and make judgments about policies, programs, and other objects in evaluation" (p. 179).

Evaluation, then, is a basic form of human behavior. Sometimes it is thorough, structured, and formal. More often it is impressionistic and private. Our focus is on the more formal, structured, and public evaluation. We want to inform readers of various approaches and methods for developing criteria and collecting information about alternatives. For those readers who aspire to become professional evaluators, we will be introducing you to the approaches and methods used in these formal studies. For all readers, practitioners and evaluators, we hope to cultivate that critical intelligence, to make you cognizant of the factors influencing your more informal judgments and decisions.

A Brief Definition of Evaluation and Other Key Terms

In the previous section, the perceptive reader will have noticed that the term "evaluation" has been used rather broadly without definition beyond what was implicit in context. But the rest of this chapter could be rather confusing if we did not stop briefly to define the term more precisely. Intuitively, it may not seem difficult to define evaluation. For example, one typical dictionary definition of evaluation is "to determine or fix the value of: to examine and judge." Seems quite straightforward, doesn't it? Yet among professional evaluators, there is no uniformly agreed-upon definition of precisely what the term "evaluation" means. In fact, in considering the role of language in evaluation, Michael Scriven, one of the founders of evaluation, for an essay on the use of language in evaluation recently noted there are nearly 60 different terms for evaluation that apply to one context or another. These include adjudge, appraise, analyze, assess, critique, examine, grade, inspect, judge, rate, rank, review, score, study, test, and so on (cited in Patton, 2000, p. 7). While all these terms may appear confusing, Scriven notes that the variety of uses of the term evaluation "reflects not only the immense importance of the process of evaluation in practical life, but the explosion of a new area of study" (cited in Patton, 2000, p. 7). This chapter will introduce the reader

to the array of variations in application, but, at this point, we will focus on one definition that encompasses many others.

Early in the development of the field, Scriven (1967) defined evaluation as judging the worth or merit of something. Many recent definitions encompass this original definition of the term (Mark, Henry, & Julnes, 2000; Schwandt, 2008; Scriven, 1991a; Stake, 2000a; Stufflebeam, 2001b). We concur that evaluation is determining the worth or merit of an evaluation object (whatever is evaluated). More broadly, we define evaluation as the identification, clarification, and application of defensible criteria to determine an evaluation object's value (worth or merit) in relation to those criteria. Note that this definition requires identifying and clarifying defensible criteria. Often, in practice, our judgments of evaluation objects differ because we have failed to identify and clarify the means that we, as individuals, use to judge an object. One educator may value a reading curriculum because of the love it instills for reading; another may disparage the program because it does not move the child along as rapidly as other curricula in helping the student to recognize and interpret letters, words, or meaning. These educators differ in the value they assign to the curricula because their criteria differ. One important role of an evaluator is to help stakeholders articulate their criteria and to stimulate dialogue about them. Our definition, then, emphasizes using those criteria to judge the merit or worth of the product.

Evaluation uses inquiry and judgment methods, including: (1) determining the criteria and standards for judging quality and deciding whether those standards should be relative or absolute, (2) collecting relevant information, and (3) applying the standards to determine value, quality, utility, effectiveness, or significance. It leads to recommendations intended to optimize the evaluation object in relation to its intended purpose(s) or to help stakeholders determine whether the evaluation object is worthy of adoption, continuation, or expansion.

Programs, Policies, and Products

In the United States, we often use the term "program evaluation." In Europe and some other countries, however, evaluators often use the term "policy evaluation." This book is concerned with the evaluation of programs, policies, and products. We are *not*, however, concerned with evaluating personnel or the performance of individual people or employees. That is a different area, one more concerned with management and personnel.[1] (See Joint Committee. [1988]) But, at this point, it would be useful to briefly discuss what we mean by programs, policies, and products. "Program" is a term that can be defined in many ways. In its simplest sense, a program is a "standing arrangement that provides for a . . . service" (Cronbach et al., 1980, p. 14). The Joint Committee on Standards for Educational Evaluation (1994) defined *program* simply as "activities that are provided on a continuing basis" (p. 3). In their

[1]The Joint Committee on Standards for Educational Evaluation has developed some standards for personnel evaluation that may be of interest to readers involved in evaluating the performance of teachers or other employees working in educational settings. These can be found at http://www.eval.org/evaluationdocuments/perseval.html.

new edition of the *Standards* (2010) the Joint Committee noted that a program is much more than a set of activities. They write:

> Defined completely, a *program* is
>
> - A set of planned systematic activities
> - Using managed resources
> - To achieve specified goals
> - Related to specific needs
> - Of specific, identified, participating human individuals or groups
> - In specific contexts
> - Resulting in documentable outputs, outcomes and impacts
> - Following assumed (explicit or implicit) systems of beliefs (diagnostic, causal, intervention, and implementation theories about how the program works)
> - With specific, investigable costs and benefits. (Joint Committee, 2010, in press)

Note that their newer definition emphasizes programs achieving goals related to particular needs and the fact that programs are based on certain theories or assumptions. We will talk more about this later when we discuss program theory. We will simply summarize by saying that a program is an ongoing, planned intervention that seeks to achieve some particular outcome(s), in response to some perceived educational, social, or commercial problem. It typically includes a complex of people, organization, management, and resources to deliver the intervention or services.

In contrast, the word "policy" generally refers to a broader act of a public organization or a branch of government. Organizations have policies—policies about recruiting and hiring employees, policies about compensation, policies concerning interactions with media and the clients or customers served by the organization. But, government bodies—legislatures, departments, executives, and others—also pass or develop policies. It might be a law or a regulation. Evaluators often conduct studies to judge the effectiveness of those policies just as they conduct studies to evaluate programs. Sometimes, the line between a program and a policy is quite blurred. Like a program, a policy is designed to achieve some outcome or change, but, unlike a program, a policy does not provide a service or activity. Instead, it provides guidelines, regulations, or the like to achieve a change. Those who study public policy define *policy* even more broadly: "public policy is the sum of government activities, whether acting directly or through agents, as it has an influence on the life of citizens" (Peters, 1999, p. 4). Policy analysts study the effectiveness of public policies just as evaluators study the effectiveness of government programs. Sometimes, their work overlaps. What one person calls a policy, another might call a program. In practice, in the United States, policy analysts tend to be trained in political science and economics, and evaluators tend to be trained in psychology, sociology, education, and public administration. As the field of evaluation expands and clients want more information on government programs, evaluators study the effectiveness of programs and policies.

Finally, a "product" is a more concrete entity than either a policy or a program. It may be a textbook such as the one you are reading. It may be a piece of software. Scriven defines a product very broadly to refer to the output of something. Thus, a product could be a student or a person who received training, the

work of a student, or a curricula which is "the product of a research and development effort" (1991a, p. 280).

Stakeholders

Another term used frequently in evaluation is "stakeholders." Stakeholders are various individuals and groups who have a direct interest in and may be affected by the program being evaluated or the evaluation's results. In the *Encyclopedia of Evaluation,* Greene (2005) identifies four types of stakeholders:

(a) People who have authority over the program including funders, policy makers, advisory boards;
(b) People who have direct responsibility for the program including program developers, administrators, managers, and staff delivering the program;
(c) People who are the intended beneficiaries of the program, their families, and their communities; and
(d) People who are damaged or disadvantaged by the program (those who lose funding or are not served because of the program). (pp. 397–398)

Scriven (2007) has grouped stakeholders into groups based on how they are impacted by the program, and he includes more groups, often political groups, than does Greene. Thus, "upstream impactees" refer to taxpayers, political supporters, funders, and those who make policies that affect the program. "Midstream impactees," also called primary stakeholders by Alkin (1991), are program managers and staff. "Downstream impactees" are those who receive the services or products of the program.

All of these groups hold a stake in the future direction of that program even though they are sometimes unaware of their stake. Evaluators typically involve at least some stakeholders in the planning and conduct of the evaluation. Their participation can help the evaluator to better understand the program and the information needs of those who will use it.

Differences in Evaluation and Research

It is important to distinguish between evaluation and research, because these differences help us to understand the distinctive nature of evaluation. While some methods of evaluation emerged from social science research traditions, there are important distinctions between evaluation and research. One of those distinctions is purpose. Research and evaluation seek different ends. The primary purpose of research is to add to knowledge in a field, to contribute to the growth of theory. A good research study is intended to advance knowledge. While the results of an evaluation study may contribute to knowledge development (Mark, Henry, & Julnes, 2000), that is a secondary concern in evaluation. Evaluation's primary purpose is to provide useful information to those who hold a stake in whatever is being evaluated (stakeholders), often helping them to make a judgment or decision.

Research seeks conclusions; evaluation leads to judgments. Valuing is the sine qua non of evaluation. A touchstone for discriminating between an evaluator and a researcher is to ask whether the inquiry being conducted would be regarded as a failure if it produced no data on the value of the thing being studied. A researcher answering strictly as a researcher will probably say no.

These differing purposes have implications for the approaches one takes. Research is the quest for laws and the development of theory—statements of relationships among two or more variables. Thus, the purpose of research is typically to explore and establish causal relationships. Evaluation, instead, seeks to examine and describe a particular thing and, ultimately, to consider its value. Sometimes, describing that thing involves examining causal relationships; often, it does not. Whether the evaluation focuses on a causal issue depends on the information needs of the stakeholders.

This highlights another difference in evaluation and research—who sets the agenda. In research, the hypotheses to be investigated are chosen by the researcher based on the researcher's assessment of the appropriate next steps in developing theory in the discipline or field of knowledge. In evaluation, the questions to be answered are not those of the evaluator, but rather come from many sources, including those of significant stakeholders. An evaluator might suggest questions, but would never determine the focus of the study without consultation with stakeholders. Such actions, in fact, would be unethical in evaluation. Unlike research, good evaluation always involves the inclusion of stakeholders—often a wide variety of stakeholders—in the planning and conduct of the evaluation for many reasons: to ensure that the evaluation addresses the needs of stakeholders, to improve the validity of results, and to enhance use.

Another difference between evaluation and research concerns generalizability of results. Given evaluation's purpose of making judgments about a particular thing, good evaluation is quite specific to the context in which the evaluation object rests. Stakeholders are making judgments about a particular evaluation object, a program or a policy, and are not as concerned with generalizing to other settings as researchers would be. In fact, the evaluator should be concerned with the particulars of that setting, with noting them and attending to the factors that are relevant to program success or failure in that setting. (Note that the setting or context may be a large, national program with many sites, or a small program in one school.) In contrast, because the purpose of research is to add to general knowledge, the methods are often designed to maximize generalizability to many different settings.

As suggested previously, another difference between research and evaluation concerns the intended use of their results. Later in the book, we will discuss the many different types of use that may occur in evaluation, but, ultimately, evaluation is intended to have some relatively immediate impact. That impact may be on immediate decisions, on decisions in the not-too-distant future, or on perspectives that one or more stakeholder groups or stakeholders have about the object of the evaluation or evaluation itself. Whatever the impact, the evaluation is designed to be used. Good research may or may not be used right away. In fact, research that adds in important ways to some theory may not be immediately noticed, and

connections to a theory may not be made until some years after the research is conducted.[2] Nevertheless, the research stands alone as good research if it meets the standards for research in that discipline or field. If one's findings are to add to knowledge in a field, ideally, the results should transcend the particulars of time and setting.

Thus, research and evaluation differ in the standards used to judge their adequacy (Mathison, 2007). Two important criteria for judging the adequacy of research are internal validity, the study's success at establishing causality, and external validity, the study's generalizability to other settings and other times. These criteria, however, are not sufficient, or appropriate, for judging the quality of an evaluation. As noted previously, generalizability, or external validity, is less important for an evaluation because the focus is on the specific characteristics of the program or policy being evaluated. Instead, evaluations are typically judged by their *accuracy* (the extent to which the information obtained is an accurate reflection—a one-to-one correspondence—with reality), *utility* (the extent to which the results serve the practical information needs of intended users), *feasibility* (the extent to which the evaluation is realistic, prudent, diplomatic, and frugal), and *propriety* (the extent to which the evaluation is done legally and ethically, protecting the rights of those involved). These standards and a new standard concerning *evaluation accountability* were developed by the Joint Committee on Standards for Evaluation to help both users of evaluation and evaluators themselves to understand what evaluations should do (Joint Committee, 2010). (See Chapter 3 for more on the Standards.)

Researchers and evaluators also differ in the knowledge and skills required to perform their work. Researchers are trained in depth in a single discipline—their field of inquiry. This approach is appropriate because a researcher's work, in almost all cases, will remain within a single discipline or field. The methods he or she uses will remain relatively constant, as compared with the methods that evaluators use, because a researcher's focus remains on similar problems that lend themselves to certain methods of study. Evaluators, by contrast, are evaluating many different types of programs or policies and are responding to the needs of clients and stakeholders with many different information needs. Therefore, evaluators' methodological training must be broad and their focus may transcend several disciplines. Their education must help them to become sensitive to the wide range of phenomena to which they must attend if they are to properly assess the worth of a program or policy. Evaluators must be broadly familiar with a wide variety of methods and techniques so they can choose those most appropriate for the particular program and the needs of its stakeholders. In addition, evaluation has developed some of its own specific methods, such as using logic models to understand program theory and metaevaluation. Mathison writes that "evaluation as a practice shamelessly borrows from all disciplines and ways of thinking to get at both facts and values" (2007, p. 20). Her statement illustrates both the methodological breadth required of an evaluator and

[2]A notable example concerns Darwin's work on evolution. Elements of his book, *The Origin of the Species*, were rejected by scientists some years ago and are only recently being reconsidered as new research suggests that some of these elements were correct. Thus, research conducted more than 100 years ago emerges as useful because new techniques and discoveries prompt scientists to reconsider the findings.

TABLE 1.1 ***Differences in Research and Evaluation***

Factor	*Research*	*Evaluation*
Purpose	Add to knowledge in a field, develop laws and theories	Make judgments, provide information for decision making
Who sets the agenda or focus?	Researchers	Stakeholders and evaluator jointly
Generalizability of results	Important to add to theory	Less important, focus is on particulars of program or policy and context
Intended use of results	Not important	An important standard
Criteria to judge adequacy	Internal and external validity	Accuracy, utility, feasibility, propriety, evaluation accountability
Preparation of those who work in area	Depth in subject matter, fewer methodological tools and approaches	Interdisciplinary, many methodological tools, interpersonal skills

the fact that evaluators' methods must serve the purpose of valuing or establishing merit and worth, as well as establishing facts.

Finally, evaluators differ from researchers in that they must establish personal working relationships with clients. As a result, studies of the competencies required of evaluators often cite the need for training in interpersonal and communication skills (Fitzpatrick, 1994; King, Stevahn, Ghere, & Minnema, 2001; Stufflebeam & Wingate, 2005).

In summary, research and evaluation differ in their purposes and, as a result, in the roles of the evaluator and researcher in their work, their preparation, and the criteria used to judge the work. (See Table 1.1 for a summary of these differences.) These distinctions lead to many differences in the manner in which research and evaluation are conducted.

Of course, evaluation and research sometimes overlap. An evaluation study may add to our knowledge of laws or theories in a discipline. Research can inform our judgments and decisions regarding a program or policy. Yet, fundamental distinctions remain. Our earlier discussion highlights these differences to help those who are new to evaluation to see the ways in which evaluators behave differently than researchers. Evaluations may add to knowledge in a field, contribute to theory development, establish causal relationships, and provide explanations for the relationship between phenomena, but that is not its primary purpose. Its primary purpose is to assist stakeholders in making value judgments and decisions about whatever is being evaluated.

Action Research

A different type of research altogether is action research. Action research, originally conceptualized by Kurt Lewin (1946) and more recently developed by Emily Calhoun (1994, 2002), is research conducted collaboratively by professionals to

improve their practice. Such professionals might be social workers, teachers, or accountants who are using research methods and means of thinking to develop their practice. As Elliott (2005) notes, action research always has a developmental aim. Calhoun, who writes of action research in the context of education, gives examples of teachers working together to conceptualize their focus; to collect, analyze, and interpret data on the issue; and to make decisions about how to improve their practice as teachers and/or a program or curriculum they are implementing. The data collection processes may overlap with program evaluation activities, but there are key differences: Action research is conducted by professionals about their own work with a goal of improving their practice. Action research is also considered to be a strategy to change the culture of organizations to one in which professionals work collaboratively to learn, examine, and research their own practices. Thus, action research produces information akin to that in formative evaluations—information to be used for program improvement. The research is conducted by those delivering the program and, in addition to improving the element under study, has major goals concerning professional development and organizational change.

The Purposes of Evaluation

Consistent with our earlier definition of evaluation, we believe that the primary purpose of evaluation is to render judgments about the value of whatever is being evaluated. This view parallels that of Scriven (1967), who was one of the earliest to outline the purpose of formal evaluation. In his seminal paper, "The Methodology of Evaluation," he argued that evaluation has a single goal or purpose: to determine the worth or merit of whatever is evaluated. In more recent writings, Scriven has continued his emphasis on the primary purpose of evaluation being to judge the merit or worth of an object (Scriven, 1996).

Yet, as evaluation has grown and evolved, other purposes have emerged. A discussion of these purposes sheds light on the practice of evaluation in today's world. For the reader new to evaluation, these purposes illustrate the many facets of evaluation and its uses. Although we agree with Scriven's historical emphasis on the purpose of evaluation, to judge the merit or worth of a program, policy, process, or product, we see these other purposes of evaluation at play as well.

Some years ago, Talmage (1982) argued that an important purpose of evaluation was "to assist decision makers responsible for making policy" (p. 594). And, in fact, providing information that will improve the quality of decisions made by policymakers continues to be a major purpose of program evaluation. Indeed, the rationale given for collecting much evaluation data today—by schools, by state and local governments, by the federal government, and by nonprofit organizations—is to help policymakers in these organizations make decisions about whether to continue programs, to initiate new programs, or, in other major ways, to change the funding or structure of a program. In addition to decisions made by policymakers, evaluation is intended to inform the decisions of many others, including program managers (principals, department heads), program staff (teachers, counselors,

health care providers, and others delivering the services offered by a program), and program consumers (clients, parents, citizens). A group of teachers may use evaluations of student performance to make decisions on program curricula or materials. Parents make decisions concerning where to send their children to school based on information on school performance. Students choose institutions of higher education based on evaluative information. The evaluative information or data provided may or may not be the most useful for making a particular decision, but, nevertheless, evaluation clearly serves this purpose.

For many years, evaluation has been used for program improvement. As we will discuss later in this chapter, Michael Scriven long ago identified program improvement as one of the roles of evaluation, though he saw that role being achieved through the initial purpose of judging merit and worth. Today, many see organizational and program improvement as a major, direct purpose of evaluation (Mark, Henry, & Julnes, 2000; Patton, 2008a; Preskill & Torres, 1998).

Program managers or those who deliver a program can make changes to improve the program based on the evaluation results. In fact, this is one of the most frequent uses of evaluation. There are many such examples: teachers using the results of student assessments to revise their curricula or pedagogical methods, health care providers using evaluations of patients' use of medication to revise their means of communicating with patients about dosage and use, and trainers using feedback from trainees to change training to improve its application on the job. These are all ways that evaluation serves the purpose of program improvement.

Today, many evaluators see evaluation being used for program and organizational improvement in new ways. As we will describe in later chapters, Michael Patton often works today in what he calls "developmental evaluation," working to assist organizations that do not have specific, measurable goals, but, instead, need evaluation to help them with ongoing progress, adaptation, and learning (Patton, 1994, 2005b). Hallie Preskill (Preskill, 2008; Preskill & Torres, 2000) and others (King, 2002; Baker & Bruner, 2006) have written about the role of evaluation in improving overall organizational performance by instilling new ways of thinking. In itself, the process of participating in an evaluation can begin to influence the ways that those who work in the organization approach problems. For example, an evaluation that involves employees in developing a logic model for the program to be evaluated or in examining data to draw some conclusions about program progress may prompt those employees to use such procedures or these ways of approaching a problem in the future and, thus, lead to organizational improvement.

The purpose of program or organizational improvement, of course, overlaps with others. When an evaluation is designed for program improvement, the evaluator must consider the decisions that those managing and delivering the program will make in using the study's results for program improvement. So the purpose of the evaluation is to provide both decision making and program improvement. We will not split hairs to distinguish between the two purposes, but will simply acknowledge that evaluation can serve both purposes. Our goal is to expand your view of the various purposes for evaluation and to help you consider the purpose in your own situation or organization.

Some recent discussions of the purposes of evaluation move beyond these more immediate purposes to evaluation's ultimate impact on society. Some evaluators point out that one important purpose of evaluation is helping give voice to groups who are not often heard in policy making or planning programs. Thus, House and Howe (1999) argue that the goal of evaluation is to foster deliberative democracy. They encourage the evaluator to work to help less powerful stakeholders gain a voice and to stimulate dialogue among stakeholders in a democratic fashion. Others highlight the role of the evaluator in helping bring about greater social justice and equality. Greene, for example, notes that values inevitably influence the practice of evaluation and, therefore, evaluators can never remain neutral. Instead, they should recognize the diversity of values that emerge and arise in an evaluation and work to achieve desirable values of social justice and equity (Greene, 2006).

Carol Weiss (1998b) and Gary Henry (2000) have argued that the purpose of evaluation is to bring about social betterment. Mark, Henry, and Julnes (2000) define achieving social betterment as "the alleviation of social problems, meeting of human needs" (p. 190). And, in fact, evaluation's purpose of social betterment is at least partly reflected in the Guiding Principles, or ethical code, adopted by the American Evaluation Association. One of those principles concerns the evaluator's responsibilities for the general and public welfare. Specifically, Principle E5 states the following:

> Evaluators have obligations that encompass the public interest and good. Because the public interest and good are rarely the same as the interests of any particular group (including those of the client or funder) evaluators will usually have to go beyond analysis of particular stakeholder interests and consider the welfare of society as a whole. (American Evaluation Association, 2004)

This principle has been the subject of more discussion among evaluators than other principles, and deservedly so. Nevertheless, it illustrates one important purpose of evaluation. Evaluations are concerned with programs and policies that are intended to improve society. Their results provide information on the choices that policymakers, program managers, and others make in regard to these programs. As a result, evaluators must be concerned with their purposes in achieving the social betterment of society. Writing in 1997 about the coming twenty-first century, Chelimsky and Shadish emphasized the global perspective of evaluation in achieving social betterment, extending evaluation's context in the new century to worldwide challenges. These include new technologies, demographic imbalances across nations, environmental protection, sustainable development, terrorism, human rights, and other issues that extend beyond one program or even one country (Chelimsky & Shadish, 1997).

Finally, many evaluators continue to acknowledge the purpose of evaluation in extending knowledge (Donaldson, 2007; Mark, Henry, & Julnes, 2000). Although adding to knowledge is the primary purpose of research, evaluation studies can add to our knowledge of social science theories and laws. They provide an opportunity to test theories in real-world settings or to test existing theories or laws with new groups by examining whether those theories hold true in new

settings with different groups. Programs or policies are often, though certainly not always, based on some theory or social science principles.[3] Evaluations provide the opportunity to test those theories. Evaluations collect many kinds of information that can add to our knowledge: information describing client groups or problems, information on causes or consequences of problems, tests of theories concerning impact. For example, Debra Rog conducted an evaluation of a large intervention program to help homeless families in the early 1990s (Rog, 1994; Rog, Holupka, McCombs-Thornton, Brito, & Hambrick, 1997). At the time, not much was known about homeless families and some of the initial assumptions in planning were incorrect. Rog adapted her evaluation design to learn more about the circumstances of homeless families. Her results helped to better plan the program, but also added to our knowledge about homeless families, their health needs, and their circumstances. In our discussion of the differences between research and evaluation, we emphasized that the primary purpose of research is to add to knowledge in a field and that this is not the primary purpose of evaluation. We continue to maintain that distinction. However, the results of some evaluations can add to our knowledge of social science theories and laws. This is not a primary purpose, but simply one purpose that an evaluation may serve.

In closing, we see that evaluation serves many different purposes. Its primary purpose is to determine merit or worth, but it serves many other valuable purposes as well. These include assisting in decision making; improving programs, organizations, and society as a whole; enhancing democracy by giving voice to those with less power; and adding to our base of knowledge.

Roles and Activities of Professional Evaluators

Evaluators as practitioners play numerous roles and conduct multiple activities in performing evaluation. Just as discussions on the purposes of evaluation help us to better understand what we mean by determining merit and worth, a brief discussion of the roles and activities pursued by evaluators will acquaint the reader with the full scope of activities that professionals in the field pursue.

A major role of the evaluator that many in the field emphasize and discuss is that of encouraging the use of evaluation results (Patton, 2008a; Shadish, 1994). While the means for encouraging use and the anticipated type of use may differ, considering use of results is a major role of the evaluator. In Chapter 17, we will discuss the different types of use that have been identified for evaluation and various means for increasing that use. Henry (2000), however, has cautioned that focusing primarily on use can lead to evaluations focused solely on program and organizational improvement and, ultimately, avoiding final decisions about merit and worth. His concern is appropriate; however, if the audience for the evaluation

[3]The term "evidence-based practice" emerges from the view that programs should be designed around social science research findings when basic research, applied research, or evaluation studies have found that a given program practice or action leads to the desired, intended outcomes.

is one that is making decisions about the program's merit and worth, this problem may be avoided. (See discussion of formative and summative evaluation in this chapter.) Use is certainly central to evaluation, as demonstrated by the prominent role it plays in the professional standards and codes of evaluation. (See Chapter 3.)

Others' discussions of the role of the evaluator illuminate the ways in which evaluators might interact with stakeholders and other users. Rallis and Rossman (2000) see the role of the evaluator as that of a critical friend. They view the primary purpose of evaluation as learning and argue that, for learning to occur, the evaluator has to be a trusted person, "someone the emperor knows and can listen to. She is more friend than judge, although she is not afraid to offer judgments" (p. 83). Schwandt (2001a) describes the evaluator in the role of a teacher, helping practitioners develop critical judgment. Patton (2008a) envisions evaluators in many different roles including facilitator, collaborator, teacher, management consultant, organizational development (OD) specialist, and social-change agent. These roles reflect his approach to working with organizations to bring about developmental change. Preskill and Torres (1998) stress the role of the evaluator in bringing about organizational learning and instilling a learning environment. Mertens (1999), Chelimsky (1998), and Greene (1997) emphasize the important role of including stakeholders, who often have been ignored by evaluation. House and Howe (1999) argue that a critical role of the evaluator is stimulating dialogue among various groups. The evaluator does not merely report information, or provide it to a limited or designated key stakeholder who may be most likely to use the information, but instead stimulates dialogue, often bringing in disenfranchised groups to encourage democratic decision making.

Evaluators also have a role in program planning. Bickman (2002), Chen (1990), and Donaldson (2007) emphasize the important role that evaluators play in helping articulate program theories or logic models. Wholey (1996) argues that a critical role for evaluators in performance measurement is helping policymakers and managers select the performance dimensions to be measured as well as the tools to use in measuring those dimensions.

Certainly, too, evaluators can play the role of the scientific expert. As Lipsey (2000) notes, practitioners want and often need evaluators with the "expertise to track things down, systematically observe and measure them, and compare, analyze, and interpret with a good faith attempt at objectivity" (p. 222). Evaluation emerged from social science research. While we will describe the growth and emergence of new approaches and paradigms, and the role of evaluators in educating users to our purposes, stakeholders typically contract with evaluators to provide technical or "scientific" expertise and/or an outside "objective" opinion. Evaluators can occasionally play an important role in making program stakeholders aware of research on other similar programs. Sometimes, the people managing or operating programs or the people making legislative or policy decisions on programs are so busy fulfilling their primary responsibilities that they are not aware of other programs or agencies that are doing similar things and the research conducted on these activities. Evaluators, who typically explore existing research on similar programs to identify potential designs and measures, can play the role

of scientific expert in making stakeholders aware of research. (See, for example, Fitzpatrick and Bledsoe [2007] for a discussion of Bledsoe's role in informing stakeholders of existing research on other programs.)

Thus, the evaluator takes on many roles. In noting the tension between advocacy and neutrality, Weiss (1998b) writes that the role(s) evaluators play will depend heavily on the context of the evaluation. The evaluator may serve as a teacher or critical friend in an evaluation designed to improve the early stages of a new reading program. The evaluator may act as a facilitator or collaborator with a community group appointed to explore solutions to problems of unemployment in the region. In conducting an evaluation on the employability of new immigrant groups in a state, the evaluator may act to stimulate dialogue among immigrants, policymakers, and nonimmigrant groups competing for employment. Finally, the evaluator may serve as an outside expert in designing and conducting a study for Congress on the effectiveness of annual testing in improving student learning.

In carrying out these roles, evaluators undertake many activities. These include negotiating with stakeholder groups to define the purpose of evaluation, developing contracts, hiring and overseeing staff, managing budgets, identifying disenfranchised or underrepresented groups, working with advisory panels, collecting and analyzing and interpreting qualitative and quantitative information, communicating frequently with various stakeholders to seek input into the evaluation and to report results, writing reports, considering effective ways to disseminate information, meeting with the press and other representatives to report on progress and results, and recruiting others to evaluate the evaluation (metaevaluation). These, and many other activities, constitute the work of evaluators. Today, in many organizations, that work might be conducted by people who are formally trained and educated as evaluators, attend professional conferences and read widely in the field, and identify their professional role as an evaluator, or by staff who have many other responsibilities—some managerial, some working directly with students or clients—but with some evaluation tasks thrown into the mix. Each of these will assume some of the roles described previously and will conduct many of the tasks listed.

Uses and Objects of Evaluation

At this point, it might be useful to describe some of the ways in which evaluation can be used. An exhaustive list would be prohibitive, filling the rest of this book and more. Here we provide only a few representative examples of uses made of evaluation in selected sectors of society.

Examples of Evaluation Use in Education

1. To empower teachers to have more say in how school budgets are allocated
2. To judge the quality of school curricula in specific content areas
3. To accredit schools that meet or exceed minimum accreditation standards

4. To determine the value of a middle school's block scheduling
5. To satisfy an external funding agency's demands for reports on effectiveness of school programs it supports
6. To assist parents and students in selecting schools in a district with school choice
7. To help teachers improve their reading program to encourage more voluntary reading

Examples of Evaluation Use in Other Public and Nonprofit Sectors

1. To decide whether to expand an urban transit program and where it should be expanded
2. To establish the value of a job training program
3. To decide whether to modify a low-cost housing project's rental policies
4. To improve a recruitment program for blood donors
5. To determine the impact of a prison's early-release program on recidivism
6. To gauge community reaction to proposed fire-burning restrictions to improve air quality
7. To determine the effect of an outreach program on the immunization of infants and children

Examples of Evaluation Use in Business and Industry

1. To improve a commercial product
2. To judge the effectiveness of a corporate training program on teamwork
3. To determine the effect of a new flextime policy on productivity, recruitment, and retention
4. To identify the contributions of specific programs to corporate profits
5. To determine the public's perception of a corporation's environmental image
6. To recommend ways to improve retention among younger employees
7. To study the quality of performance appraisal feedback

One additional comment about the use of evaluation in business and industry may be warranted. Evaluators unfamiliar with the private sector are sometimes unaware that personnel evaluation is not the only use made of evaluation in business and industry settings. Perhaps that is because the term "evaluation" has been absent from the descriptors for many corporate activities and programs that, when examined, are decidedly evaluative. Activities labeled as quality assurance, quality control, research and development, Total Quality Management (TQM), or Continuous Quality Improvement (CQI) turn out, on closer inspection, to possess many characteristics of program evaluation.

Uses of Evaluation Are Generally Applicable

As should be obvious by now, evaluation methods are clearly portable from one arena to another. The use of evaluation may remain constant, but the entity it is applied to—that is, the object of the evaluation—may vary widely. Thus, evaluation

may be used to improve a commercial product, a community training program, or a school district's student assessment system. It could be used to build organizational capacity in the Xerox Corporation, the E. F. Lilly Foundation, the Minnesota Department of Education, or the Utah Division of Family Services. Evaluation can be used to empower parents in the San Juan County Migrant Education Program, workers in the U.S. Postal Service, employees of Barclays Bank of England, or residents in east Los Angeles. Evaluation can be used to provide information for decisions about programs in vocational education centers, community mental health clinics, university medical schools, or county cooperative extension offices. Such examples could be multiplied ad infinitum, but these should suffice to make our point.

In some instances, so many evaluations are conducted of the same type of object that it prompts suggestions for techniques found to be particularly helpful in evaluating something of that particular type. An example would be Kirkpatrick's (1977; 1983; 2006) model for evaluating training efforts. In several areas, concern about how to evaluate broad categories of objects effectively has led to the development of various subareas within the field of evaluation, such as product evaluation, personnel evaluation, program evaluation, policy evaluation, and performance evaluation.

Some Basic Types of Evaluation

Formative and Summative Evaluation

Scriven (1967) first distinguished between the formative and summative roles of evaluation. Since then, the terms have become almost universally accepted in the field. In practice, distinctions between these two types of evaluation may blur somewhat, but the terms serve an important function in highlighting the types of decisions or choices that evaluation can serve. The terms, in fact, contrast two different types of actions that stakeholders might take as a result of evaluation.

An evaluation is considered to be formative if the primary purpose is to provide information for program improvement. Often, such evaluations provide information to judge the merit or worth of one part of a program. Three examples follow:

1. Planning personnel in the central office of Perrymount School District have been asked by the school board to plan a new, and later, school day for the local high schools. This is based on research showing that adolescents' biological clocks cause them to be more groggy in the early morning hours and on parental concerns about teenagers being released from school as early as 2:30 P.M. A formative evaluation will collect information (surveys, interviews, focus groups) from parents, teachers and school staff, and students regarding their views on the current school schedule calendar and ways to change and improve it. The planning staff will visit other schools using different schedules to observe these schedules and to interview school staff on their perceived effects. The planning staff will then give the information to the Late Schedule Advisory Group, which will make final recommendations for changing the existing schedule.

2. Staff with supervisory responsibilities at the Akron County Human Resources Department have been trained in a new method for conducting performance appraisals. One of the purposes of the training is to improve the performance appraisal interview so that employees receiving the appraisal feel motivated to improve their performance. The trainers would like to know if the information they are providing on conducting interviews is being used by those supervisors who complete the program. They plan to use the results to revise this portion of the training program. A formative evaluation might include observing supervisors conducting actual, or mock, interviews, as well as interviewing or conducting focus groups with both supervisors who have been trained and employees who have been receiving feedback. Feedback for the formative evaluation might also be collected from participants in the training through a reaction survey delivered either at the conclusion of the training or a few weeks after the training ends, when trainees have had a chance to practice the interview.

3. A mentoring program has been developed and implemented to help new teachers in the classroom. New teachers are assigned a mentor, a senior teacher who will provide them with individualized assistance on issues ranging from discipline to time management. The focus of the program is on helping mentors learn more about the problems new teachers are encountering and helping them find solutions. Because the program is so individualized, the assistant principal responsible for overseeing the program is concerned with learning whether it is being implemented as planned. Are mentors developing a trusting relationship with the new teachers and learning about the problems they encounter? What are the typical problems encountered? The array of problems? For what types of problems are mentors less likely to be able to provide effective assistance? Interviews, logs or diaries, and observations of meetings between new teachers and their mentors will be used to collect data to address these issues. The assistant principal will use the results to consider how to better train and lead the mentors.

In contrast to formative evaluations, which focus on program improvement, summative evaluations are concerned with providing information to serve decisions or assist in making judgments about program adoption, continuation, or expansion. They assist with judgments about a program's overall worth or merit in relation to important criteria. Scriven (1991a) has defined summative evaluation as "evaluation done for, or by, any observers or decision makers (by contrast with developers) who need valuative conclusions for any other reasons besides development" (p. 20). Robert Stake has memorably described the distinction between the two in this way: "When the cook tastes the soup, that's formative evaluation; when the guest tastes it, that's summative evaluation" (cited by Scriven, 1991a, p. 19). In the following examples we extend the earlier formative evaluations into summative evaluations.

1. After the new schedule is developed and implemented, a summative evaluation might be conducted to determine whether the schedule should be continued and expanded to other high schools in the district. The school board might be

the primary audience for this information because it is typically in a position to make the judgments concerning continuation and expansion or termination, but others—central office administrators, principals, parents, students, and the public at large—might be interested stakeholders as well. The study might collect information on attendance, grades, and participation in after-school activities. Other unintended side effects might be examined, such as the impact of the schedule on delinquency, opportunities for students to work after school, and other afternoon activities.

2. To determine whether the performance appraisal program should be continued, the director of the Human Resource Department and his staff might ask for an evaluation of the impact of the new performance appraisal on job satisfaction and performance. Surveys of employees and existing records on performance might serve as key methods of data collection.

3. Now that the mentoring program for new teachers has been tinkered with for a couple of years using the results of the formative evaluation, the principal wants to know whether the program should be continued. The summative evaluation will focus on turnover, satisfaction, and performance of new teachers.

Note that the audiences for formative and summative evaluation are very different. In formative evaluation, the audience is generally the people delivering the program or those close to it. In our examples, they were those responsible for developing the new schedule, delivering the training program, or managing the mentoring program. Because formative evaluations are designed to improve programs, it is critical that the primary audience be people who are in a position to make changes in the program and its day-to-day operations. Summative evaluation audiences include potential consumers (students, teachers, employees, managers, or officials in agencies that could adopt the program), funding sources, and supervisors and other officials, as well as program personnel. The audiences for summative evaluations are often policymakers or administrators, but can, in fact, be any audience with the ability to make a "go–no go" decision. Teachers make such decisions with curricula. Consumers (clients, parents, and students) make decisions about whether to participate in a program based on summative information or their judgments about the overall merit or worth of a program.

A Balance between Formative and Summative. It should be apparent that both formative and summative evaluation are essential because decisions are needed during the developmental stages of a program to improve and strengthen it, and again, when it has stabilized, to judge its final worth or determine its future. Unfortunately, some organizations focus too much of their work on summative evaluations. This trend is noted in the emphases of many funders today on impact or outcome assessment from the beginning of a program or policy. An undue emphasis on summative evaluation can be unfortunate because the development process, without formative evaluation, is incomplete and inefficient. Consider the foolishness of developing a new aircraft design and submitting it to a summative

test flight without first testing it in the formative wind tunnel. Program test flights can be expensive, too, especially when we haven't a clue about the probability of success.

Formative data collected during the early stages of a program can help identify problems in the program model or theory or in the early delivery of the program that can then be modified or corrected. People delivering the program may need more training or resources to effectively implement the model. The model may have to be adapted because the students or clients being served are not exactly as program developers anticipated. Perhaps they have different learning strategies or less knowledge, skills, or motivation than anticipated; therefore, the training program or class curriculum should be expanded or changed. In other cases, students or clients who participate in a program may have more, or different, skills or problems than program planners anticipated. The program, then, must be adapted to address those.[4] So, a formative evaluation can be very useful at the beginning of a program to help it succeed in achieving its intended outcomes.

Conversely, some organizations may avoid summative evaluations. Evaluating for improvement is critical, but, ultimately, many products and programs should be judged for their overall merit and worth. Henry (2000) has noted that evaluation's emphasis on encouraging use of results can lead us to serving incremental, often formative, decisions and may steer us away from the primary purpose of evaluation—determining merit and worth.

Although formative evaluations more often occur in the early stages of a program's development and summative evaluations more often occur in its later stages, it would be an error to think they are limited to those time frames. Well-established programs can benefit from formative evaluations. Some new programs are so problematic that summative decisions are made to discontinue. However, the relative emphasis on formative and summative evaluation changes throughout the life of a program, as suggested in Figure 1.1, although this generalized concept obviously may not precisely fit the evolution of any particular program.

An effort to distinguish between formative and summative evaluation on several dimensions appears in Table 1.2. As with most conceptual distinctions, formative and summative evaluation are often not as easy to distinguish in the real world as they seem in these pages. Scriven (1991a) has acknowledged that the two are often profoundly intertwined. For example, if a program continues beyond a summative evaluation study, the results of that study may be used for both summative and, later, formative evaluation purposes. In practice, the line between formative and summative is often rather fuzzy.

[4]See the interview with Stewart Donaldson about his evaluation of a work-training program (Fitzpatrick & Donaldson, 2002) in which he discusses his evaluation of a program that had been successful in Michigan, but was not adapted to the circumstances of California sites, which differed in the reasons why people were struggling with returning to the workforce. The program was designed anticipating that clients would have problems that these clients did not have.

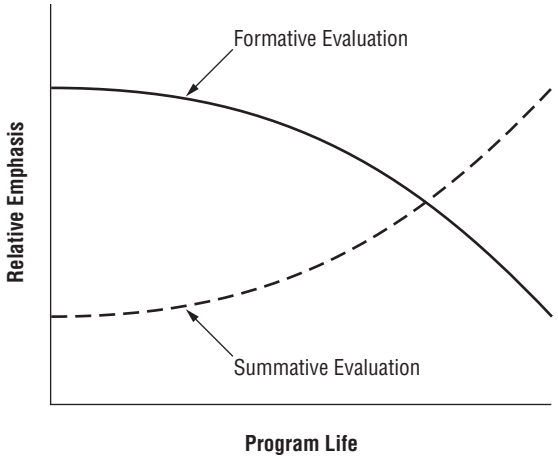


FIGURE 1.1 ***Relationship between Formative and Summative Evaluation***

TABLE 1.2 ***Differences between Formative and Summative Evaluation***

	Formative Evaluation	*Summative Evaluation*
Use	To improve the program	To make decisions about the program's future or adoption
Audience	Program managers and staff	Administrators, policymakers, and/or potential consumers or funding agencies
By Whom	Often internal evaluators supported by external evaluators	Often external evaluators, supported by internal evaluators
Major Characteristics	Provides feedback so program personnel can improve it	Provides information to enable decision makers to decide whether to continue it, or consumers to adopt it
Design Constraints	What information is needed? When?	What standards or criteria will be used to make decisions?
Purpose of Data Collection	Diagnostic	Judgmental
Frequency of Data Collection	Frequent	Infrequent
Sample Size	Often small	Usually large
Questions Asked	What is working? What needs to be improved? How can it be improved?	What results occur? With whom? Under what conditions? With what training? At what cost?

Beyond Formative and Summative. Our discussion of the purposes of evaluation reflects the changes and expansions that have occurred in the practice of evaluation over the decades. Michael Patton (1996) has described three purposes of evaluation that do not fall within the formative or summative dimension. These include the following:

1. The contribution of evaluation to conceptual thinking, rather than immediate or instrumental decisions or judgments, about an object. As evaluation practice has expanded and research has been conducted on how evaluation is used, evaluators have found that evaluation results are often not used immediately, but, rather, are used gradually—conceptually—to change stakeholders' thinking about the clients or students they serve, about the logic models or theories for programs, or about the ways desired outcomes can be achieved.

2. Evaluation for broad, long-term organizational learning and continuous improvement. Patton's developmental evaluation falls within this category. Results from such evaluations are not used for direct program improvement (formative purposes), but to help organizations consider future directions, changes, and adaptations that should be made because of new research findings or changes in the context of the program and its environment. (See Preskill [2008]; Preskill and Torres [2000].)

3. Evaluations in which the process of the evaluation may have more import than the use of the results. As we will discuss in Chapter 17, research on the use of evaluation has found that participation in the evaluation process itself, not just the results of the evaluation, can have important impacts. Such participation can change the way people plan programs in the future by providing them with skills in developing logic models for programs or by empowering them to participate in program planning and development in different ways. As we discussed, one purpose of evaluation is to improve democracy. Some evaluations empower the public or disenfranchised stakeholder groups to participate further in decision making by providing them with information or giving them a voice through the evaluation to make their needs or circumstances known to policymakers.

The distinction between formative and summative evaluations remains a primary one when considering the types of decisions the evaluation will serve. However, it is important to remember the other purposes of evaluation and, in so doing, to recognize and consider these purposes when planning an evaluation so that each evaluation may reach its full potential.

Needs Assessment, Process, and Outcome Evaluations

The distinctions between formative and summative evaluation are concerned primarily with the kinds of decisions or judgments to be made with the evaluation results. The distinction between the relative emphasis on formative or summative evaluation is an important one to make at the beginning of a study because it

informs the evaluator about the context, intention, and potential use of the study and has implications for the most appropriate audiences for the study. However, the terms do not dictate the nature of the questions the study will address. Chen (1996) has proposed a typology to permit consideration of process and outcome along with the formative and summative dimension. We will discuss that typology here, adding needs assessment to the mix.

Some evaluators use the terms "needs assessment," "process," and "outcome" to refer to the types of questions the evaluation study will address or the focus of the evaluation. These terms also help make the reader aware of the full array of issues that evaluators examine. Needs assessment questions are concerned with (a) establishing whether a problem or need exists and describing that problem, and (b) making recommendations for ways to reduce the problem; that is, the potential effectiveness of various interventions. Process, or monitoring studies, typically describe how the program is delivered. Such studies may focus on whether the program is being delivered according to some delineated plan or model or may be more open-ended, simply describing the nature of delivery and the successes and problems encountered. Process studies can examine a variety of different issues, including characteristics of the clients or students served, qualifications of the deliverers of the program, characteristics of the delivery environment (equipment, printed materials, physical plant, and other elements of the context of delivery), or the actual nature of the activities themselves. Outcome or impact studies are concerned with describing, exploring, or determining changes that occur in program recipients, secondary audiences (families of recipients, coworkers, etc.), or communities as a result of a program. These outcomes can range from immediate impacts or outputs (for example, achieving immediate learning objectives in a lesson or course) to longer-term objectives, final goals, and unintended outcomes.

Note that these terms do not have implications for how the information will be used. The terms formative and summative help us distinguish between the ways in which the results of the evaluation may be used for immediate decision making. Needs assessment, process, and outcome evaluations refer to the nature of the issues or questions that will be examined. In the past, people have occasionally misused the term formative to be synonymous with process evaluation, and summative to be synonymous with outcome evaluation. However, Scriven (1996) himself notes that "formative evaluations are not a species of process evaluation. Conversely, summative evaluation may be largely or entirely process evaluation" (p. 152).

Table 1.3 illustrates the application of these evaluation terms building on a typology proposed by Chen (1996); we add needs assessment to Chen's typology. As Table 1.3 illustrates, an evaluation can be characterized by the action the evaluation will serve (formative or summative) as well as by the nature of the issues it will address.

To illustrate, a needs assessment study can be summative (Should we adopt this new program or not?) or formative (How should we modify this program to deliver it in our school or agency?). A process study often serves formative purposes, providing information to program providers or managers about how to change activities to improve the quality of the program delivery to make it more likely that

TABLE 1.3 ***A Typology of Evaluation Studies***

		Judgment	
		What to Revise/Change Formative	*What to Begin, Continue, Expand Summative*
Focus of Questions	*Needs Assessment*	How should we adapt the model we are considering?	Should we begin a program? Is there sufficient need?
	Process	Is more training of staff needed to deliver the program appropriately?	Are sufficient numbers of the target audience participating in the program to merit continuation?
	Outcome	How can we revise our curricula to better achieve desired outcomes?	Is this program achieving its goals to a sufficient degree that its funding should be continued?

objectives will be achieved, but a process study may also serve summative purposes. A process study may reveal that the program is too complex or expensive to deliver or that program recipients (students, trainees, clients) do not enroll as expected. In such cases, a process study that began as a formative evaluation for program improvement may lead to a summative decision to discontinue the program. Accountability studies often make use of process data to make summative decisions.

An outcome study can, and often does, serve formative or summative purposes. Formative purposes may be best served by examining more immediate outcomes because program deliverers have greater control over the actions leading to these outcomes. For example, teachers and trainers often make use of immediate measures of student learning to make changes in their curriculum or methods. They may decide to spend more time on certain areas or to expand on the types of exercises or problems students practice to better achieve certain learning goals, or they may spend less time on areas in which students have already achieved competency. Policymakers making summative decisions, however, are often more concerned with the program's success at achieving other, more global outcomes, such as graduation rates or employment placement, because their responsibility is with these outcomes. Their decisions regarding funding concern whether programs achieve these ultimate outcomes. The fact that a study examines program outcomes, or effects, however, tells us nothing about whether the study serves formative or summative purposes.

Internal and External Evaluations

The adjectives "internal" and "external" distinguish between evaluations conducted by program employees and those conducted by outsiders. An experimental year-round education program in the San Francisco public schools might be evaluated by a member of the school district staff (internal) or by a site-visit team appointed by the California State Board of Education (external). A large health care organization with facilities in six communities might have a member of each facility's staff evaluate the

effectiveness of their outreach program in improving immunization rates for infants and children (internal), or the organization may hire a consulting firm or university research group to look at all six programs (external).

Seems pretty simple, right? Often it is, but how internal is the evaluation of the year-round school program if it is conducted by an evaluation unit at the central office, which is quite removed from the charter school implementing the program? Is that an internal or external evaluation? Actually, the correct answer is both, for such an evaluation is clearly external from the perspective of those in the charter school, yet might be considered an internal evaluation from the perspective of the state board of education or parents in the district.

There are obvious advantages and disadvantages connected with both internal and external evaluation roles. Table 1.4 summarizes some of these. Internal evaluators are likely to know more about the program, its history, its staff, its clients, and its struggles than any outsider. They also know more about the organization and its culture and styles of decision making. They are familiar with the kinds of information and arguments that are persuasive, and know who is likely to take action and who is likely to be persuasive to others. These very advantages, however, are also disadvantages. They may be so close to the program that they cannot see it clearly. (Note, though, that each evaluator, internal and external, will bring his or her own history and biases to the evaluation, but the internal evaluators' closeness may prevent them from seeing solutions or changes that those newer to the situation might see more readily.) While successful internal evaluators may overcome the hurdle of perspective, it can be much more difficult for them to overcome the barrier of position. If internal evaluators are not provided with sufficient decision-making power, autonomy, and protection, their evaluation will be hindered.

The strengths of external evaluators lie in their distance from the program and, if the right evaluators are hired, their expertise. External evaluators are perceived as more credible by the public and, often, by policymakers. In fact, external evaluators typically do have greater administrative and financial independence. Nevertheless, the objectivity of the external evaluator can be overestimated. (Note the role of the external Arthur Andersen firm in the 2002 Enron bankruptcy and scandal. The lure of obtaining or keeping a large contract can prompt external parties to bend the rules to

TABLE 1.4 ***Advantages of Internal and External Evaluators***

Internal	*External*
More familiar with organization & program history	Can bring greater credibility, perceived objectivity
Knows decision-making style of organization	Typically brings more breadth and depth of technical expertise for a particular evaluation
Is present to remind others of results now and in future	Has knowledge of how other similar organizations and programs work
Can communicate technical results more frequently and clearly	

keep the contract.) However, for programs with high visibility or cost or those surrounded by much controversy, an external evaluator can provide a desirable degree of autonomy from the program. External evaluators, if the search and hiring process are conducted appropriately, can also bring the specialized skills needed for a particular project. In all but very large organizations, internal evaluators must be jacks-of-all-trades to permit them to address the ongoing evaluation needs of the organization. When seeking an external evaluator, however, an organization can pinpoint and seek the types of skills and expertise needed for that time-limited project.

Organizing Internal Evaluation for Maximum Effect. In recent years, evaluations conducted by people employed by the organization have grown exponentially as funders' demands for accountability have increased. This growth is at least partly due to professional evaluators' emphasis on building internal organizational capacity to conduct evaluation. (Capacity building and mainstreaming evaluation were the conference themes for the American Evaluation Association in 2000 and 2001, respectively, with the 2001 conference focusing on one of our co-authors' themes, mainstreaming evaluation. See Leviton, [2001] and Sanders, [2002] for their published Presidential Addresses on the subjects.) We will discuss capacity building further in Chapter 9, but in this section we will discuss ways in which to structure internal evaluation to improve evaluation and the performance of the organization.

First, a comment on internal evaluators. For many years, large school districts had, and many continue to have, internal evaluation units. The economic constraints on education have reduced the number of districts with strong internal evaluation units, but such units remain in many districts. (See, for example, Christie's interview with Eric Barela, an internal evaluator with the Los Angeles Unified School District, Christie and Barela [2008]). In many nonprofit organizations, internal evaluation capacity has increased in recent years. This growth has been spurred by United Way of America (UWA), a major funding source for many nonprofit, human service organizations, which encouraged these organizations to implement its evaluation strategy for measuring outcomes (Hendricks, Plantz, & Pritchard, 2008). Today, approximately 19,000 local agencies funded by United Way conduct internal evaluations, supplemented with training by United Way, to measure agency outcomes. Similarly, Cooperative Extensions and other organizations are active in conducting internal evaluations (Lambur, 2008). State and local governments have been thrust into a more active evaluation role through federal performance-based management systems. All these efforts have prompted public and nonprofit organizations to train existing staff to, at minimum, report data on program outcomes and, often, to conduct evaluations to document those outcomes.

Given the growth in internal evaluation, it is appropriate to consider how internal evaluations can be conducted for the maximum effect. Evaluators have been writing about ways to enhance internal evaluation for some years (Chelimsky, 1994; Love, 1983, 1991; Scriven, 1975; Sonnichsen, 1987, 1999; Stufflebeam, 2002a). Probably the two most important conditions identified for successful internal evaluations are (a) active support for evaluation from top administrators within the organization and (b) clearly defined roles for internal

evaluators. The strength of internal evaluators is their ongoing contribution to decision making within the organization. Without the active support of leaders within the organization, internal evaluators cannot fulfill that role.

Where evaluators should be located in a large organization is an area of some disagreement. Internal evaluators must be situated where they can understand organizational problems, initiate or plan evaluations to address those problems, and be in a position to frequently communicate results to the stakeholders who can use them. Some argue that internal evaluators should, therefore, be placed centrally within the organization where they can work closely with top decision makers. In this way, the internal evaluators can serve in an advisory function to top managers and are able to communicate information from a variety of evaluation studies as needed. Many, if not most, internal evaluation units are centrally located in the organization and, hence, have the potential to serve in that capacity. With proximity to top managers, the director of an internal evaluation unit can continue to demonstrate the value of evaluation to the organization.

Others (Lambur, 2008), however, have argued that internal evaluators should be dispersed among program units where they can provide useful, formative evaluation for program improvement directly to people who are delivering the organization's programs. In such positions, internal evaluators can build a more trusting relationship with program deliverers and increase the chances that the results of their evaluations will be used. Lambur, in interviews with internal evaluators in cooperative extension offices, found disadvantages to being "closely aligned with administration" (2008, p. 49). Staff who are delivering programs, such as teachers, social workers, trainers, and others, see evaluation in the central office as being more concerned with accountability and responding to federal government demands and less concerned with improving programs. Lambur found evaluators who worked in program units were able to become closer to the programs, and, as a result, they believed, knew how to conduct more useful evaluations. They recognized the potential for being less objective, but worked to make their evaluations more rigorous. In such positions, internal evaluators can serve in Rallis and Rossman's role of critical friend (2000).

Patton (2008b) has also interviewed internal evaluators and has found that they face many challenges. They can be excluded from major decisions and asked to spend time on public relations functions rather than true evaluation. In addition, they do, in fact, spend much time gathering data for accountability requirements from external funding sources; this takes away time from developing relationships with administrators and people who deliver the program. Internal evaluators are often, but not always, full-time evaluators. Like many professionals in organizations, they can have other responsibilities that conflict with their evaluation role.

Patton (2008b) and Lambur (2008) argue that internal evaluators face competing demands in evaluating for accountability and for program improvement. Both argue that the emphasis for internal evaluators should be on program improvement. Lambur writes,

> "Through my personal experience [as an internal evaluator], I learned it was far more effective to promote evaluation as a tool for improving programs than helping the

organization meet demands for accountability. If program staff view themselves as primary stakeholders for evaluation results, they are more apt to become engaged in the process of conducting high-quality evaluations. Results of such evaluations can be used first for program improvement, and then for accountability purposes."

Those writing about the organization of internal evaluation acknowledge the difficulties an internal evaluator faces, but provide many useful suggestions. The solution, however, for an individual organization can depend on its mission and purpose. In some organizations, placing evaluation in a central location with top administrators can provide the distance from programs needed for credibility in important summative evaluations and can supply evaluators with avenues for affecting organizational learning and culture through educating key administrators about the role of evaluation. In other organizations, it can be important to place evaluators in program units where they can focus on the improvement of individual programs. In either case, internal evaluators require organizational support from top managers, mid-level managers, and supervisors. Internal evaluators can help create a true learning organization, where evaluation is looked to for valuable information to make decisions. To do so, though, requires careful planning and continuous communication and support from others in clarifying and supporting the role of evaluation in the organization.

Possible Role Combinations. Given the growth in internal evaluation capacity, considering how to combine internal and external evaluation is important. One way is to consider the purposes of evaluation. The dimensions of formative and summative evaluation can be combined with the dimensions of internal and external evaluation to form the two-by-two matrix shown in Figure 1.2. The most common roles in evaluation might be indicated by cells 1 and 4 in the matrix. Formative evaluations are often conducted by internal evaluators, and there are clear merits in such an approach. Their knowledge of the program, its history, staff, and clients is of great value, and credibility is not nearly the problem it would be in a summative evaluation. Program personnel are often the primary audience, and the evaluator's ongoing relationship with them can enhance the use of results in a good learning organization. Summative evaluations are probably best conducted

	Internal	**External**
Formative	1 Internal Formative	2 External Formative
Summative	3 Internal Summative	4 External Summative

FIGURE 1.2 ***Combination of Evaluation Roles***

by external evaluators. It is difficult, for example, to know how much credibility to attach to a Ford Motor Company evaluation that concludes that a particular Ford automobile is far better than its competitors in the same price range. The credibility accorded to an internal summative program evaluation (cell 3) in a school or nonprofit organization may be no better.

In some cases, though, funds are not available for external evaluators, or competent external evaluators cannot be identified. In many cases, summative evaluations are conducted internally and, in such cases, role combinations are possible to improve the credibility of the results. Patton (2008a) suggests using external evaluators to review and comment on the quality of internal evaluations. In other cases, external evaluators can design critical elements of the evaluation, helping define the evaluation questions and developing evaluation designs and measures, perhaps working jointly with an internal evaluation team. Internal evaluators can then work to implement the evaluation and to develop effective means for communicating results to different stakeholder groups. Such role combinations can save critical fiscal resources, improve internal capacity, and enhance the credibility of the results. (See, for example, Fitzpatrick's interview with Debra Rog concerning her role as an external evaluator in a project for homeless families spanning several cities. She discusses the role of staff within each organization in helping conduct and plan the evaluation with her guidance [Fitzpatrick and Rog, 1999]). In any case, when a summative evaluation is conducted internally, managers within the organization need to attend to the position of the evaluators in the organization relative to the program being evaluated. They must work to ensure maximum independence and must not place evaluators in the untenable position of evaluating programs developed by their boss or colleagues.

Sonnichsen (1999) writes of the high impact that internal evaluation can have if the organization has established conditions that permit the internal evaluator to operate effectively. The factors that he cites as being associated with evaluation offices that have a strong impact on the organization include operating as an independent entity, reporting to a top official, giving high rank to the head of the office, having the authority to self-initiate evaluations, making recommendations and monitoring their implementation, and disseminating results widely throughout the organization. He envisions the promise of internal evaluation, writing, "The practice of internal evaluation can serve as the basis for organizational learning, detecting and solving problems, acting as a self-correcting mechanism by stimulating debate and reflection among organizational actors, and seeking alternative solutions to persistent problems" (Sonnichsen, 1999, p. 78).

Evaluation's Importance—and Its Limitations

Given its many uses, it may seem almost axiomatic to assert that evaluation is not only valuable but essential in any effective system or society. Citizens look to evaluation for accountability. Policymakers and decision makers call on it and use it

to make important decisions. Program staff can use evaluation to plan and improve programs to better meet clients' and societal needs and to make decisions about how to stay within their budget. Consumers, such as parents, students, and voluntary clients, can make choices about schools for themselves or their children or the hospital, clinic, or agency they will contact for services. Evaluators can perform many roles for those delivering programs. These include helping them develop good programs, helping them deliver the programs to changing clients for changing contexts, and helping them find interventions that are most successful in achieving their goals. Evaluators can help organizations as a whole through stimulating a learning culture, thereby helping those in the organization to question and consider their goals and their methods, their clients and their needs, and showing them how to use evaluative inquiry methods to meet their needs. As some evaluators note, evaluation plays an important continuing role in democracy. It informs citizens and, thus, empowers them to influence their schools, their government, and their nonprofit organizations. It can influence the power of stakeholders who have been absent from important decisions by giving them voice and power through evaluation. Scriven (1991b) said it well:

> The process of disciplined evaluation permeates all areas of thought and practice. . . . It is found in scholarly book reviews, in engineering's quality control procedures, in the Socratic dialogues, in serious social and moral criticism, in mathematics, and in the opinions handed down by appellate courts. . . . It is the process whose duty is the systematic and objective determination of merit, worth, or value. Without such a process, there is no way to distinguish the worthwhile from the worthless. (p. 4)

Scriven also argues the importance of evaluation in pragmatic terms ("bad products and services cost lives and health, destroy the quality of life, and waste the resources of those who cannot afford waste"), ethical terms ("evaluation is a key tool in the service of justice"), social and business terms ("evaluation directs effort where it is most needed, and endorses the 'new and better way' when it is better than the traditional way—and the traditional way where it's better than the new high-tech way"), intellectual terms ("it refines the tools of thought"), and personal terms ("it provides the only basis for justifiable self-esteem") (p. 43). Perhaps for these reasons, evaluation has increasingly been used as an instrument to pursue goals of organizations and agencies at local, regional, national, and international levels.

But, evaluation's importance is not limited to the methods used, the stakeholder supplied with information, or the judgment of merit or worth that is made. Evaluation gives us a process to improve our ways of thinking and, therefore, our ways of developing, implementing, and changing programs and policies. Schwandt has argued that evaluators need to cultivate in themselves and others an intelligent belief in evaluation. He writes that "possessing (and acting on) an intelligent belief in evaluation is a special obligation of evaluators—those who claim to be well prepared in the science and art of making distinctions of worth" (2008, p. 139). He reminds us that evaluation is not simply the methods, or tools, that we use, but a way of thinking. Citing some problematic trends we see in society today, the political

manipulation of science and the tendency to see or argue for all-or-nothing solutions that must be used in all settings in the same way, Schwandt calls for evaluators to help citizens and stakeholders use better means of reasoning. This better means of reasoning would draw on the kinds of thinking good evaluators should do. The characteristics of such reasoning include a tolerance for ambiguity, a recognition of multiple perspectives and a desire to learn from those different perspectives, a desire to experiment or to become what Don Campbell called an "experimenting society." Describing this society and evaluation's role in it, Schwandt writes:

> This is a society in which we ask serious and important questions about what kind of society we should have and what directions we should take. This is a social environment indelibly marked by uncertainty, ambiguity, and interpretability. Evaluation in such an environment is a kind of social conscience; it involves serious questioning of social direction; and it is a risky undertaking in which we endeavor to find out not simply whether what we are doing is a good thing but also what we do not know about what we are doing. So we experiment—we see what we can learn from different ways of knowing. In evaluation, we try to work from the top down (so to speak) using what policy makers say they are trying to do as a guide, as well as from the bottom up, doing evaluation that is heavily participant oriented or user involved. All this unfolds in an atmosphere of questioning, of multiple visions of what it is good to do, of multiple interpretations of whether we as a society are doing the right thing. (2008, p. 143)

As others in evaluation have done, Schwandt is reminding us of what evaluation should be. As evaluators, we learn how to use research methods from many disciplines to provide information and reach judgments about programs and policies, but our methods and theories underlie an approach to reasoning. This approach is its greatest promise.

Limitations of Evaluation. In addition to its potential for impact, evaluation has many limitations. Although the purpose of this book is to help the reader learn how to conduct good evaluations, we would be remiss if we did not discuss these limitations. The methods of evaluation are not perfect ones. No single study, even those using multiple methods, can provide a wholly accurate picture of the truth because truth is composed of multiple perspectives. Formal evaluation is more successful than informal evaluation, in part, because it is more cautious and more systematic. Formal evaluation is guided by explicit questions and criteria. It considers multiple perspectives. Its methods allow one to follow the chain of reasoning, the evaluative argument, and to more carefully consider the accuracy, or the validity, of the results. But evaluations are constrained by realities, including some characteristics of the program and its context, the competencies of the evaluation staff, the budget, the timeframe, and the limits of what measures can tell us.

A more important limitation to evaluation than the methodological and fiscal ones, however, are the political ones. We live in a democracy. That means that elected, and appointed, officials must attend to many issues. Results of evaluations

are not their sole source of information by any means, nor should they be. Citizens' input and expectations obviously play a role in decisions. Many stakeholder groups, experts, lawmakers, policymakers, and, yes, lobbyists, have information and experience that are important to consider. So, in the best of situations, evaluation is simply one piece of information, albeit an important piece, we hope, in the marble cake of sources used by decision makers in a democracy.

Finally, both evaluators and their clients may have been limited by a tendency to view evaluation as a series of discrete studies rather than a continuing system representing an approach to reasoning and personal and organizational growth. It can be difficult to question what you do and the activities that you believe in, but evaluative inquiry must prompt us to do that, both in evaluating our evaluations (metaevaluation) and in evaluating programs. A few poorly planned, badly executed, or inappropriately ignored evaluations should not surprise us; such failings occur in every field of human endeavor. This book is intended to help evaluators, and the policymakers, managers, and all the other stakeholders who participate in and use evaluations, to improve their evaluative means of reasoning and to improve the practice of evaluation.

Major Concepts and Theories

1. Evaluation is the identification, clarification, and application of defensible criteria to determine an evaluation object's value, its merit or worth, in regard to those criteria. The specification and use of explicit criteria distinguish formal evaluation from the informal evaluations most of us make daily.

2. Evaluation differs from research in its purpose, the role of the evaluator and the researcher in determining the focus of the study, the criteria used to judge its quality, its involvement of stakeholders, and the competencies required of those who practice it.

3. The basic purpose of evaluation is to render judgments about the value of the object under evaluation. Other purposes include providing information for program and organizational improvement and to make decisions, working to better society and to improve and sustain democratic values, encouraging meaningful dialogue among many diverse stakeholders, as well as adding to our knowledge concerning the application of social science theory, and providing oversight and compliance for programs.

4. Evaluators play many roles including facilitator, planner, advocate, scientific expert, critical friend, collaborator, and aid to decision makers and other stakeholder groups.

5. Evaluations can serve formative or summative decisions as well as other purposes. Formative evaluations are designed for program improvement. The audience is, most typically, stakeholders close to the program. Summative evaluations serve decisions about program adoption, continuation, or expansion. Audiences for these evaluations must have the ability to make such "go-no go" decisions.

6. Evaluations can address needs assessment, process, or outcome questions. Any of these types of questions can serve formative or summative purposes.

7. Evaluators may be internal or external to the organization. Internal evaluators know the organizational environment and can facilitate communication and use of results. External evaluators can provide more credibility in high-profile evaluations and bring a fresh perspective and different skills to the evaluation.

8. Evaluation goes beyond particular methods and tools to include a way of thinking. Evaluators have a role in educating stakeholders and the public about the concept of evaluation as a way of thinking and reasoning. This way of thinking includes acknowledging, valuing, using, and exploring different perspectives and ways of knowing, and creating and encouraging an experimenting society—one that actively questions, considers, and creates policies, programs, interventions, and ideas.

Discussion Questions

1. Consider a program in your organization. If it were to be evaluated, what might be the purpose of the evaluation at this point in time? Consider the stage of the program and the information needs of different stakeholder groups. What might be the role of evaluators in conducting the evaluation?
2. What kind of evaluation do you think is most useful—formative or summative? What kind of evaluation would be most useful to you in your work? To your school board or elected officials?
3. Which do you prefer, an external or internal evaluator? Why?
4. Describe a situation in which an internal evaluator would be more appropriate than an external evaluator. What is the rationale for your choice? Now describe a situation in which an external evaluator would be more appropriate.

Application Exercises

1. List the types of evaluation studies that have been conducted in an institution or agency of your acquaintance, noting in each instance whether the evaluator was internal or external to that institution. Determine whether each study was formative or summative and whether it was focused on needs assessment, process, or outcome questions. Did the evaluation address the appropriate questions? If not, what other types of questions or purposes might it have addressed?
2. Think back to any formal evaluation study you have seen conducted (or if you have never seen one conducted, find a written evaluation report of one). Identify three things that make it different from informal evaluations. Then list ten informal evaluations you have performed so far today. (Oh, yes you have!)
3. Discuss the potential and limitations of program evaluation. Identify some things evaluation can and cannot do for programs in your field.
4. Within your own organization (if you are a university student, you might choose your university), identify several evaluation objects that you believe would be appropriate for study. For each, identify (a) the stakeholder groups and purposes the evaluation study would serve, and (b) the types of questions the evaluation might address.

Case Studies

In this edition, we begin a new practice to acquaint readers with real evaluations in order to give them a better understanding of the practice of evaluation. At the end of many chapters, we will recommend one or more interviews that Jody Fitzpatrick, one of our authors, or Christina Christie conducted with a well-known evaluator concerning one evaluation he or she completed. Each article begins with a brief summary of the evaluation. Fitzpatrick or Christie then interviews the evaluator about the choices he or she made in determining the purposes of the evaluation, involving stakeholders, selecting designs and data collection methods, collecting the data, reporting the results, and facilitating use. Interested readers may refer to the book that collects and analyzes these interviews:

Fitzpatrick, J. L., Christie, C. A., & Mark, M. M. (2008). *Evaluation in action: Interviews with expert evaluators*. Thousand Oaks, CA: Sage.

Or, the reader may read individual interviews published in the *American Journal of Evaluation*.

For this chapter we recommend two interviews to orient the reader to two quite different types of evaluation in *Evaluation in Action* Chapters 1 (James Riccio) and 7 (Gary Henry).

In Chapter 1, James Riccio describes the choices he made in an evaluation designed to judge the merit and worth of a welfare reform program for the state of California as welfare reform initiatives first began. His major stakeholder is the California legislature, and the study illustrates a traditional, mixed-methods evaluation with significant instrumental use. The journal source is as follows: Fitzpatrick, J. L. & Riccio, J. (1997). A dialogue about an award-winning evaluation of GAIN: A welfare-to-work program. *Evaluation Practice, 18*, 241–252.

In Chapter 7, Gary Henry describes the development of a school "report card" for schools in Georgia during the early stages of the performance monitoring emphasis for K–12 education. The evaluation provides descriptive information to help parents, citizens, and policymakers in Georgia learn more about the performance of individual schools. The journal source is as follows: Fitzpatrick, J. L., & Henry, G. (2000). The Georgia Council for School Performance and its performance monitoring system: A dialogue with Gary Henry. *American Journal of Evaluation, 21,* 105–117.

Suggested Readings

Greene, J. C. (2006). Evaluation, democracy, and social change. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation.* London: Sage Publications.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). Toward an integrative framework for evaluation practice. *American Journal of Evaluation, 20*, 177–198.

Patton, M. Q. (1996). A world larger than formative and summative. *Evaluation Practice, 17*(2), 131–144.

Rallis, S. F., & Rossman, G. B. (2000). Dialogue for learning: Evaluator as critical friend. In R. K. Hopson (Ed.), *How and why language matters in evaluation.* New Directions for Evaluation, No. 86, 81–92. San Francisco: Jossey-Bass.

Schwandt, T. A. (2008). Educating for intelligent belief in evaluation. *American Journal of Evaluation, 29*(2), 139–150.

Sonnichsen, R. C. (1999). *High impact internal evaluation.* Thousand Oaks, CA: Sage.

Stake, R. E. (2000). A modest commitment to the promotion of democracy. In K. E. Ryan & L. DeStefano (Eds.), *Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation.* New Directions for Evaluation, No. 85, 97–106. San Francisco: Jossey-Bass.

2

Origins and Current Trends in Modern Program Evaluation

Orienting Questions

1. How did the early stages of evaluation influence practice today?
2. What major political events occurred in the late 1950s and early 1960s that greatly accelerated the growth of evaluation thought?
3. What significant events precipitated the emergence of modern program evaluation?
4. How did evaluation evolve as a profession in the 1970s and 1980s?
5. How has evaluation changed in the last two decades? What factors have influenced these changes?

Formal evaluation of educational, social, and private-sector programs is still maturing as a field, with its most rapid development occurring during the past four decades. Compared with professions such as law, education, and accounting or disciplines like sociology, political science, and psychology, evaluation is still quite new. In this chapter, we will review the history of evaluation and its progress toward becoming a full-fledged profession and transdiscipline. This history and the concluding discussion of the current state of evaluation will make the reader better aware of all the directions that evaluation can take.

The History and Influence of Evaluation in Society

Early Forms of Formal Evaluation

Some evaluator-humorists have mused that formal evaluation was probably at work in determining which evasion skills taught in Sabertooth Avoidance 101 had the

greatest survival value. Scriven (1991c) apparently was not speaking tongue-in-cheek when suggesting that formal evaluation of crafts may reach back to the evaluation of early stone-chippers' products, and he was obviously serious in asserting that it can be traced back to samurai sword evaluation.

In the public sector, formal evaluation was evident as early as 2000 B.C., when Chinese officials conducted civil service examinations to measure the proficiency of applicants for government positions. And in education, Socrates used verbally mediated evaluations as part of the learning process. But centuries passed before formal evaluations began to compete with religious and political beliefs as the driving force behind social and educational decisions.

Some commentators see the ascendancy of natural science in the seventeenth century as a necessary precursor to the premium that later came to be placed on direct observation. Occasional tabulations of mortality, health, and populations grew into a fledgling tradition of empirical social research that grew until "In 1797, Encyclopedia Britannica could speak of statistics—'state-istics,' as it were—as a 'word lately introduced to express a view or survey of any kingdom, county, or parish'" (Cronbach et al., 1980, p. 24).

But quantitative surveys were not the only precursor to modern social research in the 1700s. Rossi and Freeman (1985) give an example of an early British sea captain who divided his crew into a "treatment group" that was forced to consume limes, and a "control group" that consumed the sailors' normal diet. Not only did the experiment show that "consuming limes could avert scurvy," but "British seamen eventually were forced to consume citrus fruits—this is the derivation of the label 'limeys,' which is still sometimes applied to the English" (pp. 20–21).

Program Evaluation: 1800–1940

During the 1800s, dissatisfaction with educational and social programs in Great Britain generated reform movements in which government-appointed royal commissions heard testimony and used other less formal methods to evaluate the respective institutions. This led to still-existing systems of external inspectorates for schools in England and much of Europe. Today, however, those systems use many of the modern concepts of evaluation; for example, recognition of the role of values and criteria in making judgments and the importance of context. Inspectorates visit schools to make judgments concerning quality and to provide feedback for improvement. Judgments may be made about the quality of the school as a whole or the quality of teachers, subjects, or themes. (See Standaert. [2000])

In the United States, educational evaluation during the 1800s took a slightly different bent, being influenced by Horace Mann's comprehensive annual, empirical reports on Massachusetts's education in the 1840s and the Boston School Committee's 1845 and 1846 use of printed tests in several subjects—the first instance of wide-scale assessment of student achievement serving as the basis for school comparisons. These two developments in Massachusetts were the first attempts at objectively measuring student achievement to assess the quality of a large school system. They set a precedent

seen today in the standards-based education movement's use of test scores from students as the primary means for judging the effectiveness of schools.

Later, during the late 1800s, liberal reformer Joseph Rice conducted one of the first comparative studies in education designed to provide information on the quality of instructional methods. His goal was to document his claims that school time was used inefficiently. To do so, he compared a large number of schools that varied in the amount of time spent on spelling drills and then examined the students' spelling ability. He found negligible differences in students' spelling performance among schools where students spent as much as 100 minutes a week on spelling instruction in one school and as little as 10 minutes per week in another. He used these data to flog educators into seeing the need to scrutinize their practices empirically.

The late 1800s also saw the beginning of efforts to accredit U.S. universities and secondary schools, although that movement did not really become a potent force for evaluating educational institutions until several strong regional accrediting associations were established in the 1930s. The early 1900s saw another example of accreditation (broadly defined) in Flexner's (1910) evaluation—backed by the American Medical Association and the Carnegie Foundation—of the 155 medical schools then operating in the United States and Canada. Although based only on one-day site visits to each school by himself and one colleague, Flexner argued that inferior training was immediately obvious: "A stroll through the laboratories disclosed the presence or absence of apparatus, museum specimens, library and students; and a whiff told the inside story regarding the manner in which anatomy was cultivated" (Flexner, 1960, p. 79). Flexner was not deterred by lawsuits or death threats from what the medical schools viewed as his "pitiless exposure" of their medical training practices. He delivered his evaluation findings in scathing terms. For example, he called Chicago's fifteen medical schools "the plague spot of the country in respect to medical education" (p. 84). Soon "schools collapsed to the right and left, usually without a murmur" (p. 87). No one was ever left to wonder whether Flexner's reports were evaluative.

Other areas of public interest were also subjected to evaluation in the early 1900s; Cronbach and his colleagues (1980) cite surveys of slum conditions, management and efficiency studies in the schools, and investigations of local government corruption as examples. Rossi, Freeman, and Lipsey (1998) note that evaluation first emerged in the field of public health, which was concerned with infectious diseases in urban areas, and in education, where the focus was on literacy and occupational training.

Also in the early 1900s, the educational testing movement began to gain momentum as measurement technology made rapid advances under E. L. Thorndike and his students. By 1918, objective testing was flourishing, pervading the military and private industry as well as all levels of education. The 1920s saw the rapid emergence of norm-referenced tests developed for use in measuring individual performance levels. By the mid-1930s, more than half of the United States had some form of statewide testing, and standardized, norm-referenced testing, including achievement tests and personality and interest profiles, became a huge commercial enterprise.

During this period, educators regarded measurement and evaluation as nearly synonymous, with the latter usually thought of as summarizing student test

performance and assigning grades. Although the broader concept of evaluation, as we know it today, was still embryonic, useful measurement tools for the evaluator were proliferating rapidly, even though very few meaningful, formally published evaluations of school programs or curricula would appear for another 20 years. One notable exception was the ambitious, landmark Eight Year Study (Smith & Tyler, 1942) that set a new standard for educational evaluation with its sophisticated methodology and its linkage of outcome measures to desired learning outcomes. Tyler's work, in this and subsequent studies (e.g., Tyler, 1950), also planted the seeds of standards-based testing as a viable alternative to norm-referenced testing. (We will return in Chapter 6 to the profound impact that Tyler and those who followed in his tradition have had on program evaluation, especially in education.)

Meanwhile, foundations for evaluation were being laid in fields beyond education, including human services and the private sector. In the early decades of the 1900s, Fredrick Taylor's scientific management movement influenced many. His focus was on systemization and efficiency—discovering the most efficient way to perform a task and then training all staff to perform it that way. The emergence of "efficiency experts" in industry soon permeated the business community and, as Cronbach et al. (1980) noted, "business executives sitting on the governing boards of social services pressed for greater efficiency in those services" (p. 27). Some cities and social agencies began to develop internal research units, and social scientists began to trickle into government service, where they started conducting applied social research in specific areas of public health, housing needs, and work productivity. However, these ancestral, social research "precursors to evaluation" were small, isolated activities that exerted little overall impact on the daily lives of the citizenry or the decisions of the government agencies that served them.

Then came the Great Depression and the sudden proliferation of government services and agencies as President Roosevelt's New Deal programs were implemented to salvage the U.S. economy. This was the first major growth in the federal government in the 1900s, and its impact was profound. Federal agencies were established to oversee new national programs in welfare, public works, labor management, urban development, health, education, and numerous other human service areas, and increasing numbers of social scientists went to work in these agencies. Applied social research opportunities abounded, and soon social science academics began to join with their agency-based colleagues to study a wide variety of variables relating to these programs. While some scientists called for explicit evaluation of these new social programs (e.g., Stephan, 1935), most pursued applied research at the intersection of their agency's needs and their personal interests. Thus, sociologists pursued questions that were of interest to the discipline of sociology and the agency, but the questions of interest often emerged from sociology. The same trend occurred with economists, political scientists, and other academics who came to conduct research on federal programs. Their projects were considered to be "field research" and provided opportunities to address important questions within their discipline in the field. (See the interview with Michael Patton in the "Suggested Readings" at the end of this chapter for an example. In this interview, he discusses how his dissertation was initially planned as field research in sociology but led Patton into the field of evaluation.)

Program Evaluation: 1940–1964

Applied social research expanded during World War II as researchers investigated government programs intended to help military personnel in areas such as reducing their vulnerability to propaganda, increasing morale, and improving the training and job placement of soldiers. In the following decade, studies were directed at new programs in job training, housing, family planning, and community development. As in the past, such studies often focused on particular facets of the program in which the researchers happened to be most interested. As these programs increased in scope and scale, however, social scientists began to focus their studies more directly on entire programs rather than on the parts of them they found personally intriguing.

With this broader focus came more frequent references to their work as "evaluation research" (social research methods applied to improve a particular program).[1] If we are liberal in stretching the definition of evaluation to cover most types of data collection in health and human service programs, we can safely say evaluation flourished in those areas in the 1950s and early 1960s. Rossi et al. (1998) state that it was commonplace during that period to see social scientists "engaged in evaluations of delinquency-prevention programs, felon-rehabilitation projects, psychotherapeutic and psychopharmacological treatments, public housing programs, and community organization activities" (p. 23). Such work also spread to other countries and continents. Many countries in Central America and Africa were the sites of evaluations examining health and nutrition, family planning, and rural community development. Most such studies drew on existing social research methods and did not extend the conceptual or methodological boundaries of evaluation beyond those already established for behavioral and social research. Such efforts would come later.

Developments in educational program evaluation between 1940 and 1965 were unfolding in a somewhat different pattern. The 1940s generally saw a period of consolidation of earlier evaluation developments. School personnel devoted their energies to improving standardized testing, quasi-experimental design, accreditation, and school surveys. The 1950s and early 1960s also saw considerable efforts to enhance the Tylerian approach by teaching educators how to state objectives in explicit, measurable terms and by providing taxonomies of possible educational objectives in the cognitive domain (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956) and the affective domain (Krathwohl, Bloom, & Masia, 1964).

In 1957, the Soviets' successful launch of *Sputnik I* sent tremors through the U.S. establishment that were quickly amplified into calls for more effective teaching of math and science to American students. The reaction was immediate. Passage of the National Defense Education Act (NDEA) of 1958 poured millions of dollars into massive, new curriculum development projects, especially in mathematics and science. Only a few projects were funded, but their size and perceived importance led policymakers to fund evaluations of most of them.

The resulting studies revealed the conceptual and methodological impoverishment of evaluation in that era. Inadequate designs and irrelevant reports were

[1]We do not use this term in the remainder of the book because we think it blurs the useful distinction between research and evaluation that we outlined in the previous chapter.

only some of the problems. Most of the studies depended on imported behavioral and social science research concepts and techniques that were fine for research but not very suitable for evaluation of school programs.

Theoretical work related directly to evaluation (as opposed to research) did not exist, and it quickly became apparent that the best theoretical and methodological thinking from social and behavioral research failed to provide guidance on how to carry out many aspects of evaluation. Therefore, educational scientists and practitioners were left to glean what they could from applied social, behavioral, and educational research. Their gleanings were so meager that Cronbach (1963) penned a seminal article criticizing past evaluations and calling for new directions. Although his recommendations had little immediate impact, they did catch the attention of other education scholars, helping to spark a greatly expanded conception of evaluation that would emerge in the next decade.

The Emergence of Modern Program Evaluation: 1964–1972

Although the developments discussed so far were not sufficient in themselves to create a strong and enduring evaluation movement, each helped create a context that would give birth to such a movement. Conditions were right for accelerated conceptual and methodological development in evaluation, and the catalyst was found in the War on Poverty and the Great Society, the legislative centerpieces of the administration of U.S. President Lyndon Johnson. The underlying social agenda of his administration was an effort to equalize and enhance opportunities for all citizens in virtually every sector of society. Millions of dollars were poured into programs in education, health, housing, criminal justice, unemployment, urban renewal, and many other areas.

Unlike the private sector, where accountants, management consultants, and R & D departments had long existed to provide feedback on corporate programs' productivity and profitability, these huge, new social investments had no similar mechanism in place to examine their progress. There were government employees with some relevant competence—social scientists and technical specialists in the various federal departments, particularly in the General Accounting Office (GAO)[2]—but they were too few and not sufficiently well organized to deal even marginally with determining the effectiveness of these vast government innovations. To complicate matters, many inquiry methodologies and management techniques that worked on smaller programs proved inadequate or unwieldy with programs of the size and scope of these sweeping social reforms.

For a time it appeared that another concept developed and practiced successfully in business and industry might be successfully adapted for evaluating these federal programs, the Planning, Programming, and Budgeting System (PPBS). PPBS was part of the systems approach used in the Ford Motor Company—and

[2]This was the original name of the GAO. In 2004, its name was changed to the Government Accountability Office.

later brought to the U.S. Department of Defense (DOD) by Robert McNamara when he became Kennedy's secretary of defense. The PPBS was a variant of the systems approaches that were being used by many large aerospace, communications, and automotive industries. It was aimed at improving system efficiency, effectiveness, and budget allocation decisions by defining organizational objectives and linking them to system outputs and budgets. Many thought the PPBS would be ideally suited for the federal agencies charged with administering the War on Poverty programs, but few of the bureaucrats heading those agencies were eager to embrace it., However, PPBS was a precursor to the evaluation systems the federal government has mandated in recent years with the Government Performance Results Act (GPRA) and the Program Assessment Rating Tool (PART).

PPBS, with its focus on monitoring, outputs, and outcomes, did not succeed. Instead, the beginning of modern evaluation in the United States, Canada, and Germany was inspired by a desire to improve programs through learning from experimentation on social interventions. Ray Rist, in his research with the Working Group on Policy and Program Evaluation, which was created by the International Institute on Administrative Sciences (IIAS) to study differences in evaluation across countries, placed the United States, Canada, Germany, and Sweden among what they called "first wave" countries (Rist, 1999). These were countries that began modern evaluation in the 1960s and 1970s with the goal of improving social programs and interventions. Evaluations were often part of program planning, and evaluators were located close to the programs they were evaluating. As we will discuss later in the chapter, evaluation in the early part of the twenty-first century is more akin to the earlier PPBS systems than to its first-wave origins.

The stage for serious evaluation in the United States was set by several factors. Administrators and managers in the federal government were new to managing such large programs and felt they needed help to make them work. Managers and policymakers in government and social scientists were interested in learning more about what was working. They wanted to use the energy and funds appropriated for evaluation to begin to learn how to solve social problems. Congress was concerned with holding state and local recipients of program grants accountable for expending funds as prescribed. The first efforts to add an evaluative element to any of these programs were small, consisting of congressionally-mandated evaluations of a federal juvenile delinquency program in 1962 (Weiss, 1987) and a federal manpower development and training program enacted that same year (Wholey, 1986). It matters little which was first, however, because neither had any lasting impact on the development of evaluation. Three more years would pass before Robert F. Kennedy would trigger the event that would send a shock wave through the U.S. education system, awakening both policymakers and practitioners to the importance of systematic evaluation.

The Elementary and Secondary Education Act. The one event that is most responsible for the emergence of contemporary program evaluation is the passage of the Elementary and Secondary Education Act (ESEA) of 1965. This bill proposed a huge increase in federal funding for education, with tens of thousands of federal grants to local schools, state and regional agencies, and universities. The

largest single component of the bill was Title I (later renamed Chapter 1), destined to be the most costly federal education program in U.S. history. Wholey and White (1973) called Title I the "grand-daddy of them all" among the array of legislation that influenced evaluation at the time.

When Congress began its deliberations on the proposed ESEA, concerns began to be expressed, especially on the Senate floor, that no convincing evidence existed that any federal funding for education had ever resulted in any real educational improvements. Indeed, there were some in Congress who believed federal funds allocated to education prior to ESEA had sunk like stones into the morass of educational programs with scarcely an observable ripple to mark their passage. Robert F. Kennedy was the most persuasive voice insisting that the ESEA require each grant recipient to file an evaluation report showing what had resulted from the expenditure of the federal funds. This congressional evaluation mandate was ultimately approved for Title I (compensatory education) and Title III (innovative educational projects). The requirements, while dated today, "reflected the state-of-the-art in program evaluation at that time" (Stufflebeam, Madaus, & Kellaghan, 2000, p. 13). These requirements, which reflected an astonishing amount of micromanagement at the congressional level but also the serious congressional concerns regarding accountability, included using standardized tests to demonstrate student learning and linking outcomes to learning objectives.

Growth of Evaluation in Other Areas. Similar trends can be observed in other areas as the Great Society developed programs in job training, urban development, housing, and other anti-poverty programs. Federal government spending on anti-poverty and other social programs increased by 600% after inflation from 1950 to 1979 (Bell, 1983). As in education, people wanted to know more about how these programs were working. Managers and policymakers wanted to know how to improve the programs and which strategies worked best to achieve their ambitious goals. Congress wanted information on the types of programs to continue funding. Increasingly, evaluations were mandated. In 1969, federal spending on grants and contracts for evaluation was $17 million. By 1972, it had expanded to $100 million (Shadish, Cook, & Leviton, 1991). The federal government expanded greatly to oversee the new social programs but, just as in education, the managers, political scientists, economists, and sociologists working with them were new to managing and evaluating such programs. Clearly, new evaluation approaches, methods, and strategies were needed, as well as professionals with a somewhat different training and orientation to apply them. (See interviews with Lois-Ellin Datta and Carol Weiss cited in the "Suggested Readings" at the end of this chapter to learn more about their early involvement in evaluation studies with the federal government at that time. They convey the excitement, the expectations, and the rapid learning curve required to begin this new endeavor of studying government programs to improve the programs themselves.)

Theoretical and methodological work related directly to evaluation did not exist. Evaluators were left to draw what they could from theories in cognate disciplines and to glean what they could from better-developed methodologies, such as experimental

design, psychometrics, survey research, and ethnography. In response to the need for more specific writing on evaluation, important books and articles emerged. Suchman (1967) published a text reviewing different evaluation methods and Campbell (1969b) argued for more social experimentation to examine program effectiveness. Campbell and Stanley's book (1966) on experimental and quasi-experimental designs was quite influential. Scriven (1967), Stake (1967), and Stufflebeam (1968) began to write articles about evaluation practice and theories. At the Urban Institute, Wholey and White (1973) recognized the political aspects of evaluation being conducted within organizations. Carol Weiss's influential text (1972) was published and books of evaluation readings emerged (Caro, 1971; Worthen & Sanders, 1973). Articles about evaluation began to appear with increasing frequency in professional journals. Together, these publications resulted in a number of new evaluation models to respond to the needs of specific types of evaluation (e.g., ESEA Title III evaluations or evaluations of mental health programs).

Some milestone evaluation studies that have received significant attention occurred at this time. These included not only the evaluations of Title I, but evaluations of Head Start and the television series *Sesame Street*. The evaluations of *Sesame Street* demonstrated some of the first uses of formative evaluation, as portions of the program were examined to provide feedback to program developers for improvement. The evaluations of Great Society programs and other programs in the late 1960s and early 1970s were inspired by the sense of social experimentation and the large goals of the Great Society programs. Donald Campbell, the influential research methodologist who trained quite a few leaders in evaluation, wrote of the "experimenting society" in his article "Reforms as Experiments" urging managers to use data collection and "experiments" to learn how to develop good programs (Campbell, 1969b). He argued that managers should advocate not for their program, but for a solution to the problem their program was designed to address. By advocating for solutions and the testing of them, managers could make policymakers, citizens, and other stakeholders more patient with the difficult process of developing programs to effectively reduce tough social problems such as crime, unemployment, and illiteracy. In an interview describing his postgraduate fellowship learning experiences with Don Campbell and Tom Cook, William Shadish discusses the excitement that fueled the beginning of modern evaluation at that time, noting, "There was this incredible enthusiasm and energy for social problem solving. [We wanted to know] How does social change occur and how does evaluation contribute to that?" (Shadish & Miller, 2003, p. 266).

Graduate Programs in Evaluation Emerge. The need for specialists to conduct useful evaluations was sudden and acute, and the market responded. Congress provided funding for universities to launch new graduate training programs in educational research and evaluation, including fellowship stipends for graduate study in those specializations. Several universities began graduate programs aimed at training educational or social science evaluators. In related fields, schools of public administration grew from political science to train administrators to manage and oversee government programs, and policy analysis emerged as a growing

new area. Graduate education in the social sciences ballooned. The number of people completing doctoral degrees in economics, education, political science, psychology, and sociology grew from 2,845 to 9,463, an increase of 333%, from 1960 to 1970 (Shadish et al., 1991). Many of these graduates pursued careers evaluating programs in the public and nonprofit sectors. The stage for modern program evaluation was set by the three factors we have described: a burgeoning economy in the United States after World War II, dramatic growth in the role of the federal government in education and other policy areas during the 1960s, and, finally, an increase in the number of social science graduates with interests in evaluation and policy analysis (Shadish et al., 1991).

Evaluation Becomes a Profession: 1973–1989

This period can be characterized as one of increasing development of a distinct field of evaluation through the growth in approaches, programs to train students to become evaluators, and professional associations. At the same time, the sites of evaluation began to diversify dramatically, with the federal government playing a less dominant role.

Several prominent writers in the field proposed new and differing models. Evaluation moved beyond simply measuring whether objectives were attained, as evaluators began to consider information needs of managers and unintended outcomes. Values and standards were emphasized, and the importance of making judgments about merit and worth became apparent. These new and controversial ideas spawned dialogue and debate that fed a developing evaluation vocabulary and literature. Scriven (1972), working to move evaluators beyond the rote application of objectives-based evaluation, proposed goal-free evaluation, urging evaluators to examine the processes and context of the program to find unintended outcomes. Stufflebeam (1971), responding to the need for evaluations that were more informative to decision makers, developed the CIPP model. Stake (1975b) proposed responsive evaluation, moving evaluators away from the dominance of the experimental, social science paradigms. Guba and Lincoln (1981), building on Stake's qualitative work, proposed naturalistic evaluation, leading to much debate over the relative merits of qualitative and quantitative methods. Collectively, these new conceptualizations of evaluation provided new ways of thinking about evaluation that greatly broadened earlier views, making it clear that good program evaluation encompasses much more than simple application of the skills of the empirical scientists. (These models and others will be reviewed in Part Two.)

This burgeoning body of evaluation literature revealed sharp differences in the authors' philosophical and methodological preferences. It also underscored a fact about which there was much agreement: Evaluation is a multidimensional technical and political enterprise that requires both new conceptualizations and new insights into when and how existing methodologies from other fields might be used appropriately. Shadish and his colleagues (1991) said it well when, in recognizing the need for unique theories for evaluation, they noted that "as evaluation matured, its theory took on its own special character that resulted from the interplay among

problems uncovered by practitioners, the solutions they tried, and traditions of the academic discipline of each evaluator, winnowed by 20 years of experience" (p. 31).

Publications that focused exclusively on evaluation grew dramatically in the 1970s and 1980s, including journals and series such as *Evaluation and Program Planning, Evaluation Practice, Evaluation Review, Evaluation Quarterly, Educational Evaluation and Policy Analysis, Studies in Educational Evaluation, Canadian Journal of Program Evaluation, New Directions for Program Evaluation, Evaluation and the Health Professions, ITEA Journal of Tests and Evaluation, Performance Improvement Quarterly,* and the *Evaluation Studies Review Annual*. Others that omit evaluation from the title but highlight it in their contents included *Performance Improvement Quarterly, Policy Studies Review,* and the *Journal of Policy Analysis and Management.* In the latter half of the 1970s and throughout the 1980s, the publication of evaluation books, including textbooks, reference books, and even compendia and encyclopedias of evaluation, increased markedly. In response to the demands and experience gained from practicing evaluation in the field, a unique evaluation content developed and grew.

Simultaneously, professional associations and related organizations were formed. The American Educational Research Association's Division H was an initial focus for professional activity in evaluation. During this same period, two professional associations were founded that focused exclusively on evaluation: the Evaluation Research Society (ERS) and Evaluation Network. In 1985, these organizations merged to form the American Evaluation Association. In 1975, the Joint Committee on Standards for Educational Evaluation, a coalition of 12 professional associations concerned with evaluation in education and psychology, was formed to develop standards that both evaluators and consumers could use to judge the quality of evaluations. In 1981, they published *Standards for Evaluations of Educational Programs, Projects, and Materials.* In 1982, the Evaluation Research Society developed a set of standards, or ethical guidelines, for evaluators to use in practicing evaluation (Evaluation Research Society Standards Committee, 1982). (These *Standards* and the 1995 Guiding Principles, a code of ethics developed by the American Evaluation Association to update the earlier ERS standards, will be reviewed in Chapter 3.) These activities contributed greatly to the formalization of evaluation as a profession with standards for judging the results of evaluation, ethical codes for guiding practice, and professional associations for training, learning, and exchanging ideas.

While the professional structures for evaluation were being formed, the markets for evaluation were changing dramatically. The election of Ronald Reagan in 1980 brought about a sharp decline in federal evaluations as states were given block grants, and spending decisions and choices about evaluation requirements were delegated to the states. However, the decline in evaluation at the federal level resulted in a needed diversification of evaluation, not only in settings, but also in approaches (Shadish et al., 1991). Many state and local agencies began doing their own evaluations. Foundations and other nonprofit organizations began emphasizing evaluation. As the funders of evaluation diversified, the nature and methods of evaluation adapted and changed. Formative evaluations that examine programs to provide feedback for incremental change and improvement and to find the links

between program actions and outcomes became more prominent. Michael Patton's utilization-focused evaluation, emphasizing the need to identify a likely user of the evaluation and to adapt questions and methods to that user's needs, became a model for many evaluators concerned with use (Patton, 1975, 1986). Guba and Lincoln (1981) urged evaluators to make greater use of qualitative methods to develop "thick descriptions" of programs, providing more authentic portrayals of the nature of programs in action. David Fetterman also began writing about alternative methods with his book on ethnographic methods for educational evaluation (Fetterman, 1984). Evaluators who had previously focused on policymakers (e.g., Congress, cabinet-level departments, legislators) as their primary audience began to consider multiple stakeholders and more qualitative methods as different sources funded evaluation and voiced different needs. Participatory methods for involving many different stakeholders, including those often removed from decision making, emerged and became prominent. Thus, the decline in federal funding, while dramatic and frightening for evaluation at the time, led to the development of a richer and fuller approach to determining merit and worth.

1990–The Present: History and Current Trends

Today, evaluations are conducted in many different settings using a variety of approaches and methods. Evaluation is well established as a profession and is, as LaVelle and Donaldson remark, "growing in leaps and bounds" in recent years (2010, p. 9). Many jobs are available. Although many evaluators continue to come to the profession from other disciplines, the number of university-based evaluation training programs in the United States grew from 38 in 1994 to 48 in 2008 (LaVelle and Donaldson, 2010). Almost 6,000 people belong to the American Evaluation Association (AEA) and another 1,800 belong to the Canadian Evaluation Society (CES). In 2005, the CES and AEA sponsored a joint conference in Toronto that attracted 2,300 evaluators, including many members and attendees from other countries. Policymakers and managers in government and nonprofit settings know of, and often request or require, evaluations. For many, evaluation—funding it, managing it, or conducting it—is one of their responsibilities. So, evaluators, at least those in the United States and Canada, are no longer struggling with establishing their discipline. But in the years since 1990, evaluation has faced several important changes that influence its practice today.

Spread of Evaluation to Other Countries

Evaluation has grown rapidly in other countries in recent years. This internationalization of evaluation has influenced the practice of evaluation as evaluators adapt to the context of their country and the expectations and needs of stakeholders. Today, there are more than 75 regional and national evaluation associations around the world (Preskill, 2008). Major associations include the European Evaluation Society, the Australasian Evaluation Society, the United Kingdom Evaluation Society,

and the African Evaluation Association. The International Organization for Cooperation in Evaluation (IOCE) was created in 2003 by its 24 members of national or regional evaluation associations, with a mission to "help legitimate and support evaluation associations, societies, and networks so that they can better contribute to good governance, effective decision making, and strengthening the role of civil society" (IOCE, 2003, para 3).

As noted earlier in this chapter, Ray Rist and his colleagues identified the United States, Canada, Germany, and Sweden as countries in the "first wave" of modern evaluation that began in the late 1960s and early 1970s during a period of social experimentation. Evaluation in these first-wave countries was linked to that social experimentation and to program improvement (Rist, 1990). Rist and his colleagues identified a "second wave" of European countries where evaluation started in a different context.[3] In these second-wave countries, which included the United Kingdom, the Netherlands, Denmark, and France, evaluation began as an effort to control federal budgets and reduce government spending. The focus of evaluation was more on accountability and identifying unproductive programs than on social experimentation and program improvement. Given its purposes, evaluation in these second-wave countries was often housed centrally, near those who made decisions regarding budgets and priorities. Rist and his colleagues found that the initial impetus for evaluation in a country often had a strong influence on the subsequent conduct and purposes of evaluation in that country. A more recent evaluation influence in Europe has been the European Union and the evaluation mandates of the European Commission. For many countries in Eastern Europe, responding to these evaluation mandates is their first venture into evaluation.

Evaluation in different cultures and other countries is an exciting venture, not only because evaluation can be beneficial in helping address policy questions and issues in those countries, but also because North American evaluators can learn new methods and organizational approaches from the efforts of those in other countries (Mertens, 1999). As any traveler knows, seeing and experiencing a culture different from one's own is an eye-opener to the peculiarities—both strengths and constraints—of one's own culture. Practices or mores that had not been previously questioned are brought to our attention as we observe people or institutions in other cultures behaving differently. Citizens differ in their expectations and beliefs regarding their government, its actions, and what they want and expect to know about their government.[4] Ways in which programs are judged, feedback

[3]The research of Rist and his colleagues focused only on Europe, Canada, and the United States.

[4]For example, a French evaluator, when interviewed by Fitzpatrick, commented that the mistrust that Americans have of their government creates a fertile ground for evaluation because citizens want to know what the government is doing and what mistakes it is making. He felt French citizens lacked that suspicion of government actions and, hence, were less interested in evaluation. Patton, in the interview cited at the end of the chapter, comments on cultural differences between Japan and the United States that had implications for evaluation. In his work in Japan, he observed that blaming or calling attention to mistakes is avoided and, thus, evaluation findings would be handled differently than in the United States.

is given, or participation is sought differ across cultures and countries. These differences, of course, have implications for evaluators who must pay attention to the political and cultural context of the evaluation in order to plan and implement a study that will be trusted and used. We believe the twenty-first century will be a time for evaluators in the Western world to learn from the practices of their colleagues in other countries and that these efforts will both strengthen our own work and spread the culture of evaluation—collecting data to judge programs and form decisions—around the world.

Nonevaluators Take on Internal Evaluation Responsibilities

Another change in evaluation in recent years concerns the number and types of people carrying out evaluation-related tasks. As evaluation expanded, many people—managers, supervisors, and other program professionals—began having responsibilities for evaluation as one part of their job. As noted in this history, evaluation has often been conducted by people without specific training in evaluation. Beginning in the 1960s when social science researchers began conducting evaluation studies to meet the demand, evaluation has often had to rely on those without specific education or training in evaluation to conduct studies. In earlier years, those people were often social scientists who had training in methodology and research, but were not familiar with evaluation theories and particular concerns about context and use. Social science researchers continue to conduct evaluations today. However, many learn about the discipline of evaluation and supplement their methodological expertise with further reading, training, and attendance at evaluation conferences, as the discipline of evaluation grows and becomes better known. New today are the increasing numbers of managers and program staff who lack the methodological training in evaluation and in social science research methods, but are often responsible for internal evaluations (Datta, 2006).

Evaluation in the nonprofit sector provides an excellent example of the extent to which in-house evaluators, typically program managers and staff with other program responsibilities, have also become responsible for major components of data collection and evaluation in their organizations. More than 900,000 nonprofit and religious organizations deliver the majority of social service programs in the United States (Carman, Fredericks, and Introcaso, 2008). Most of these organizations receive funds from the 1,300 local United Way organizations, and United Way requires these organizations to conduct evaluations of their programs. The United Way approach to evaluation has admirable components, including significant training, but most of the evaluations, with United Way encouragement, are conducted by existing staff with occasional guidance from external evaluators (Hendricks, Plantz, and Pritchard, 2008). Hendricks et al. (2008), who are otherwise pleased with many elements of the United Way approach, are concerned that the overreliance on current employees who lack evaluation expertise may shortchange the organizations when it comes to effective use of the results. Survey

studies of evaluators provide further evidence of the increase in numbers of evaluators who are both internal to the organization and have other responsibilities within the organization. Christie (2003) found that many of the evaluators she surveyed in California were internal and held other, generally management, responsibilities. Many had little or no training in evaluation and were unfamiliar with evaluation theories and approaches.

In education, school districts have been faced with serious budget constraints and many have coped with these fiscal constraints by cutting central office staff, including evaluation departments. Schools, faced with increasing evaluation demands in the current standards-based environment, have had to cope with these demands with fewer evaluation professionals. As a result, teachers and administrators often face additional evaluation responsibilities. The expansion of evaluation has, therefore, had some unintended consequences that have implications for building organizational capacity and for improving education and training.

Many people involved in conducting in-house evaluations have primary professional identifications other than evaluation. They are often not interested in becoming full-time evaluators and, hence, university-based education is not the best option for providing training for these individuals. (See Datta [2006] for her discussion of the need to learn more about the evaluations produced by these practitioners to consider how their training needs can be addressed.) Expanded training opportunities and creative thinking by those in the evaluation field are needed to help these people develop their evaluation skills. Evaluation kits abound, but often focus on basic methodological issues such as designing a survey, and not on critical issues such as carefully defining purpose, involving stakeholders, and considering use.

Although the explosion of employees in organizations conducting evaluation has serious implications for training and for the accuracy, credibility, and use of evaluation studies, the move to involve other employees of schools and organizations in evaluation also has great advantages. In 2000 and 2001, the conference themes of both presidents of the American Evaluation Association addressed, in different ways, the issue of working with other employees in organizations to improve evaluation quality and use. Noting the increasing demand for evaluation and, yet, evaluators' continued struggles in affecting programs and policies, Laura Leviton, the president of AEA in 2000, used her theme of "Evaluation Capacity Building" to discuss ways to build evaluators' collective capacity to conduct better evaluations. Her suggestions included recognizing and using the strengths of program practitioners in program logic and implementation, in organizational behavior, and in the people skills needed to help those within organizations understand and use evaluation (Leviton, 2001). Rather than maintain a distance from managers and others in the program being evaluated as some evaluators have done, Leviton encouraged evaluators to learn from these people with experience in the organization. James Sanders, our co-author and AEA president in 2001, chose as his theme, "Mainstreaming Evaluation." In his opening remarks, Sanders noted that when he and Blaine Worthen published the first edition of this book in 1973, they began with the observation that "evaluation is one of the most widely discussed but little used

processes in today's systems" (2002, p. 253). He notes that the status of evaluation has improved but that it is still not second nature to organizations. Explaining his concept of mainstreaming evaluation, Sanders said, "Mainstreaming refers to the process of making evaluation an integral part of an organization's everyday operations. Instead of being put aside in the margins of work, evaluation becomes a routine part of the organization's work ethic if it is mainstreamed. It is part of the culture and job responsibilities at all levels of the organization" (2002, p. 254). Today, with much attention being paid to evaluation and accountability and with many managers and other employees playing a part in conducting evaluations, we have that opportunity. As noted earlier, the spread of evaluation responsibilities has its risks, but it also has potential benefits to evaluation and to the organization. We can cope with the risks by expanding training opportunities and by making use of partnerships between internal and external evaluators, as discussed in Chapter 1. Meanwhile, the fact that many employees of organizations, schools, and other agencies who do not identify themselves as evaluators are now involved in evaluation presents an opportunity for evaluation to become part of the culture of the organization. But, this will succeed only if we proceed carefully. Just as social scientists who came to evaluation in the 1960s often erred in viewing evaluation as simply the application of research methods in the field, today's busy managers or professionals who are conducting evaluation while balancing other responsibilities in their organization may view evaluation as simply collecting some data and reporting it to others. Sanders' concept of mainstreaming evaluation includes carefully crafting the purposes of evaluation for organizational learning and use.

A Focus on Measuring Outcomes and Impact

Another major trend that emerged in evaluation during the 1990s is the emphasis on measuring outcomes and using evaluation for purposes of accountability. The United States began evaluation in what Ray Rist and his colleagues (1999) called "first wave" evaluation with a focus on innovative experimentation and collecting data to improve programs and test new interventions. However, in many ways, the United States has transformed to a "second wave" country with a focus on evaluation for accountability and, at least the claim is, for using results to make summative and budgetary decisions about program continuation and expansion. The outcomes focus began in the early 1990s and continues unabated today.

In education, the foundation for the current standards-based outcome focus began in 1983 with the publication of *A Nation at Risk* (National Commission on Excellence in Education, 1983). That report expressed serious concerns about the state of education in the United States and provided the impetus for change. The message, which continues today, was that education in the United States was broken and that the federal government needed to become more involved to fix it. The nature of that action was not determined for a few years, but gradually a federal role with a focus on accountability emerged. Historically, local school districts and, to a lesser extent, the states have been responsible for schools in the United States. Therefore, an increased federal role in an issue that had historically been based on

local community needs was somewhat controversial. However, in 1989, the National Governors Association met with then-President George H.W. Bush at the President's Educational Summit with Governors and endorsed national goals for education while still maintaining state and local control. Later, President Clinton, who had led the National Governors Association in meeting with President Bush at the 1989 summit, greatly increased both the role of the federal government in education and the emphasis on standards with six major pieces of legislation that he signed in 1994. Press releases indicted that "not since the 1960s has so much significant education legislation been enacted" and that the six acts "promise to alter the landscape of American education in important and lasting ways" (http://www.ed.gov/PressReleases/10-1994/legla.html). The legislation included the Improving America's Schools Act (IASA), an amendment to the old 1965 Elementary and Secondary School Act that had marked the beginning of modern evaluation, and the Goals 2000: Educate America Act. Among other things, these acts provided financial support and incentives for states to develop high standards for academic achievement, to guide learning, and to monitor schools' progress toward achieving these standards. By the end of 1994, 40 states had applied for planning funds to begin developing standards. The argument was that local authority would be maintained by having states develop their own standards; the federal government's role was to require standards and to provide fiscal incentives for doing so. In 2001, under President George W. Bush's leadership, Congress passed legislation that has been the focus of educational reform ever since—the No Child Left Behind (NCLB) Act. This legislation greatly increased the federal role by establishing more requirements for student performance, testing, and teacher training, and by adding fiscal sanctions and corrective action when goals were not achieved.[5] Of course, standards and methods of assessment vary greatly across the 50 states, but, in each state, standards and their means of assessment serve as the focus for educational reform and much of educational evaluation today. Lauren Resnick writes, "Test-driven accountability has become a reality [in education]" (2006, p. 33), adding that "enormous weight is placed on tests and accountability formulas (2006, p. 37)."[6]

These policies have greatly changed the role of evaluation in public schools in the United States. Standards and their assessment receive much public attention and, in most states, are a significant driver of educational policies, practices, and evaluation. Evaluation in K–12 education in the United States today focuses on several related issues: developing appropriate means for assessing students and their progress, identifying successful schools and schools that are failing, and identifying practices that can help bring students' performance up to the standards. As schools that do not meet the standards can be closed or faculty and administrators changed, the evaluation focus is both summative (Should a school continue or not? Be re-staffed or closed?) and formative (Which students in a given school are failing to meet a standard? What have

[5]President Obama has now proposed changing No Child Left Behind, but no specific legislation has yet been passed on the issue.

[6]Resnick's special issue of *Educational Measurement: Issues and Practice* focuses on case studies of four states and how standards and measures of assessment have been put into practice and used.

been their experiences? What are the experiences of similar students who succeed? What types of interventions may be most appropriate to help those students who have not met the standard?). Such evaluation efforts can, of course, improve schools, but the focus on standards, and their assessment, also holds risks. It has changed the focus of evaluation in education to standards and accountability at a time when resources are scarce and many school evaluation efforts are able to focus on little else.

Reactions and Policy Statements. In recent years, the American Evaluation Association (AEA) has taken its first policy positions on the issues of testing and education accountability. In 2000, AEA President James Sanders, our co-author, appointed a Task Force on High Stakes Testing in K–12 Education to review the research and to develop a statement of the organization's position. The AEA Position Statement on High Stakes Testing in PreK–12 Education was passed by the AEA Board in 2002 and can be found on the AEA web site at www.eval.org/hst3.htm. The statement summarizes research on the risks and benefits of high stakes testing, concluding that "evidence of the impact of high stakes testing shows it to be an evaluative practice where the harm outweighs the benefits" (2002, p. 1). The Task Force wrote:

> Although used for more than two decades, state mandated high stakes testing has not improved the quality of schools; nor diminished disparities in academic achievement along gender, race, or class lines; nor moved the country forward in moral, social, or economic terms. The American Evaluation Association (AEA) is a staunch supporter of accountability, but not test driven accountability. AEA joins many other professional associations in opposing the inappropriate use of tests to make high stakes decisions. (2002, p. 1)

The Task Force presents other avenues for improved evaluation practice, including better validation of current tests for the purposes for which they are used, use of multiple measures, and consideration of a wide range of perspectives, including those of professional teachers to assess student performance. In 2006, the AEA Board approved a second policy statement on the issue of educational accountability (see http://www.eval.org/edac.statement.asp.) This statement expresses concerns with three major issues:

> Overreliance on standardized test scores that are not necessarily accurate measures of student learning, especially for very young and for historically underserved students, and that do not capture complex educational processes or achievements;
>
> Definitions of success that require test score increases that are higher or faster than historical evidence suggests is possible; and
>
> A one-size-fits-all approach that may be insensitive to local contextual variables or to local educational efforts (American Evaluation Association, http://www.eval.org/edac.statement.asp, 2006, p. 1)

This AEA policy statement encourages use of multiple measures, measures of individual student progress over time, context-sensitive reporting, use of data to consider resource allocations for teachers and schools, accessible appeals processes, and public participation and access.

Choice in Education. Another factor influencing evaluation in the educational environment today is school choice. Choice is represented in many different ways across the country. Some cities (Washington, DC. and Milwaukee, Wisconsin being prominent examples) have had voucher and choice systems for some time and much research has been conducted on these systems (Buckley & Schneider, 2006; Goldring & Shapira, 1993; Hoxby, 2000). In many school districts, parents now are able to send their child to another public school within the district or, in some cases, outside the district. Districts across the United States have many different choice plans, from traditional neighborhood schools to magnet schools, charter schools, and, in some areas, vouchers to private schools. The choice environment in K–12 education has, of course, influenced evaluation practice. The theory of choice is based on the market theory that competition improves performance; therefore, giving parents a choice of schools will inspire schools to become more competitive, which will improve school performance and student achievement (Chubb & Moe, 1990).

In some districts, evaluation plays a role in helping educational administrators and teachers in individual schools or groups of schools to consider how they want to market their school to recruit other students. New programs emerge; old ones are put aside. At minimum, schools struggle with predicting their enrollments and planning to staff their schools adequately. In addition, school administrators and teachers work to develop and implement new programs designed to improve learning or draw more, and sometimes, better students. Such choices, which are new to public school administrators, present challenging decision demands. What programs, curricula, or interventions will improve the schools' scores on standards? What programs, curricula, or interventions will attract more students to the school? Traditional evaluation methods can, and are, used to help teachers and administrators deal with such decisions and provide opportunities for evaluation to serve new uses. For example, Fitzpatrick has been involved in studies that examine how low-income parents who are perhaps most likely to lose out in choice environments have learned about school choice and made choices for their children (Teske, Fitzpatrick, & Kaplan, 2006). These studies are designed to help school districts better inform parents about choices. In this environment, there is much that evaluators can do to help teachers and administrators adapt to change and improve learning. (See Rodosky and Munoz [2009] for an example of how one urban school district manages its evaluation responsibilities for accountability.)

Performance Monitoring in Other Governmental Sectors. Just as education was becoming concerned with standards, their assessment, and evaluation for accountability in the late 1990s and early part of this century, other government entities and nonprofit organizations also began focusing on performance monitoring and evaluating outcomes.[7] The early influences in the trend to measure outcomes in

[7]Although the history in these other arenas is a little different from that of education, the theory and approach behind the focus on outcomes in both education and other sectors are the same. Therefore, it is helpful for those in both arenas, education and agencies that deliver other services, to be aware of the similar pressures to measure outcomes and the forces that influence each.

government came from New Public Management, a movement in public administration and management, and the related call to "reinvent government." In 1992, David Osborne and Ted Gaebler authored the popular and influential book, *Reinventing Government,* which urged public policymakers and managers to build on the successes of the private sector that was then experimenting with re-engineering and Total Quality Management (TQM). Osborne and Gaebler advocated an entrepreneurial, consumer-driven government in which managers viewed citizens as "consumers" and government managers became more entrepreneurial in developing and experimenting with programs, policies, and interventions.[8] Reinventing government was not without its critics. (See, for example, deLeon and Denhardt [2000] and their concerns with how the economic-based, market model of reinventing government and viewing citizens as consumers might neglect the broader public interest.) However, reinventing government and its principles was widely implemented in many state and local governments as well as at the federal level. During the Clinton administration, Vice-President Al Gore authored the *National Performance Review*, a government report to guide change, based on Osborne and Gaebler's principles of reinvention (National Performance Review, 1993). The report and its recommendations were intended to encourage public managers to be entrepreneurial to deal with budget constraints and to become more efficient but, at the same time, to meet citizen needs.

An important part of reinventing government was, of course, accountability or collecting data to see what worked and what didn't. Therefore, the Clinton administration also proposed the Government Performance Results Act (GPRA) to address concerns about accountability with these new initiatives (Radin, 2006). (See OBM Watch [2000] http://www.ombwatch.org/node/326 for more on GPRA.) GPRA was an example of performance monitoring measurement being advocated and implemented by several countries, including Canada and Australia in the late 1990s (Perrin, 1998; Winston, 1999). Joseph Wholey, a prominent leader in evaluation in the U.S. government in the 1970s, was involved in the development of GPRA and was a leader in performance measurement (Wholey, 1996). Passed in 1994 with implementation beginning in 1997, GPRA required all federal agencies to produce a strategic plan and to measure progress toward meeting the goals and objectives delineated in the plan with performance data. Thus, GPRA was the first major federal government mandate to measure program or policy outcomes. Government employees across the country became well acquainted with GPRA and its requirements as different levels of government responded to the requirements to identify and measure outcomes.

[8]Note the similarity between the theories of reinventing government and the theories concerning school choice. Both emerge from concepts about the market and the "success" in the private sector and a belief that public institutions can become more successful by becoming more like the private sector or businesses. Managers and school principals become "entrepreneurs" and clients, parents, and students become "consumers" or "customers" who are making choices and decisions about services. Given the economic failures of the private sector seen in the United States and around the world in 2008 and 2009, we have chosen to use quotation marks around the word success because economists and citizens are now not so certain about the successes of the private sector. Entrepreneurial behavior, without regulation, appears to have prompted the housing crisis and many problems with banks and security firms.

The Bush administration continued the emphasis on performance-based management and measuring outcomes with its own measure to replace GPRA, the Program Assessment Rating Tool (PART) (OMB, 2004). PART is a 25-item questionnaire designed to obtain information on program performance. Scores are calculated for each program based on agencies' responses, and one-half of the PART score is based on results or outcomes. Each year, the Office of Management and Budget (OMB) obtains PART scores from 20% of all government programs; programs are required to complete PART on a rotating basis, so that all programs are reviewed within five years. By 2008, 98% of federal programs had completed PART forms and been reviewed. (See http://www.whitehouse.gov/omb/part/.) Just as scores on standards-based tests can influence the staffing and even the continuation of individual schools, PART scores are intended to be used to make budgetary decisions. As in education, instances of the dramatic use of PART scores to slash funding for programs are relatively rare, but the interest in outcomes and results has been clearly established (Gilmour & Davis, 2006).

Outcomes Measurement in the Nonprofit Arena. Schools and other public organizations have not been the only ones to move to an outcomes orientation in recent years. Nonprofit organizations, as well, now focus their evaluation activities on assessing and reporting outcomes. As mentioned earlier, United Way influences much of the evaluation in the nonprofit sector. Foundations and other philanthropic organizations that fund nonprofits also influence evaluations in this arena through their grant requirements. These funding agencies have encouraged nonprofit organizations to measure their outcomes. United Way's evaluation system is called the Outcomes Measurement System and, as the name suggests, the focus is on outcomes. Other elements of the system include developing logic models to link inputs, activities, outputs, and outcomes; encouraging quantitative and repeated measures of outcomes; and emphasizing use of results for program improvement. The activities are not labeled as "evaluation" by United Way but, instead, are considered "a modest effort simply to track outcomes" (Hendricks et al., 2008, p. 16). However, the activities generally take the place of traditional evaluation efforts. The United Way model has influenced the nonprofit field broadly. There are, however, a couple of noteworthy differences between the United Way model and the outcomes focus in education and other public agencies: (a) in the United Way model accountability is considered secondary to the purpose of program improvement; (b) expectations for measuring outcomes are generally more realistic than requirements for public-sector agencies. For example, the Office of Management and Budget, in discussing evidence for outcomes, strongly encourages use of Randomized Control Trials, or RCTs (OMB, 2004).[9] United Way, recognizing that many nonprofit

[9]OMB's advocacy of randomized control trials will be discussed in Chapter 15 on design. Randomized experiments are certainly one way of establishing causality, but, along with the American Evaluation Association, we believe there are many established approaches to determining causality and the one selected should be appropriate for the context of the program and the judgments and decisions to be drawn from the evaluation.

human service organizations lack resources to conduct sophisticated evaluations of all outcomes, prefers to view the process as performance monitoring of outcomes without attempts to clearly establish causality. Nonprofit organizations, like many public sector organizations, had typically reported inputs and activities to funders. The move to assess and monitor program outcomes can be a step in the right direction in providing a more comprehensive assessment of a program.

Considering Organizational Learning and Evaluation's Larger Potential Impacts

A related trend that has influenced evaluation in the early part of the twenty-first century is a discussion of the role of evaluation in organizational learning. People in many different, but related, fields—public management, adult learning, workplace learning, organizational management and change, educational administration, leadership, and evaluation—are all writing about organizational learning and looking for ways to build organizations' capacity to learn and manage in difficult times. Senge's 1990 book on the learning organization introduced many to the theories and research in this area and prompted managers, policymakers, and others to begin thinking more about how organizations learn and change. Since evaluators are concerned with getting stakeholders within organizations to use evaluation information, obviously the concept of organizational learning was important. Preskill and Torres' book, *Evaluative Inquiry for Learning in Organizations* (1998), was one of the first to bring these concepts to the attention of evaluators through their proposal for evaluative inquiry. But other evaluation theories and approaches and the experiences of evaluators in the field were also converging to prompt evaluators to think more broadly about the role of evaluation in organizations and the tasks evaluators should perform. As early as 1994, Reichardt, in an article reflecting on what we had learned from evaluation practice, suggested that evaluators should become more involved in the planning stages of programs, because the skills that evaluators brought to the table might be more useful in the beginning stages than after programs were completed. Evaluators' increasing use of logic models to identify the focus of an evaluation and to put that focus in an appropriate context made program stakeholders more aware not only of logic models, but also of evaluative modes of thinking (Rogers & Williams, 2006). Patton (1996) coined the term "process use" to refer to changes that occur in stakeholders, often program deliverers and managers, who participate in an evaluation. These changes occur not because of specific information gained from the evaluation results, but, instead, because of what they learned from participating in the evaluation process. The evaluation process itself prompts them to think in new ways in the future. This learning may include something as direct as using logic models to develop programs or being more comfortable and confident in using data to make decisions.

Thus, the concept of learning organizations, introduced from other disciplines, and evaluators' reflections and observations on their role in organizations and their potential impact converged and prompted evaluators to move beyond the traditional focus on instrumental use of results to consider broader uses of evaluation and ways to achieve those uses more effectively.

All the changes we have discussed here—standards-based movements, the focus on outcomes, and the government's and United Way's focus on employees collecting data using on going internal systems—were also designed to change the culture of organizations and to improve organizational learning and decision making. These changes have often been initiated by people outside evaluation, such as policymakers; public administrators; and people from management, budgeting, or finance. The evaluators involved in creating performance monitoring systems such as GPRA or United Way's focus are often from different schools of evaluation than those who are advocating organizational learning through empowerment evaluation or evaluative inquiry. Nevertheless, the directions of all these changes are to modify and improve organizations' ways of learning and making decisions. Some methods are likely to be more successful than others, but the overwhelming change in this period is for evaluators to begin thinking of evaluation in broader terms. In the past, evaluators and their clients have tended to see evaluations as discrete studies to be used for a particular problem or policy, rather than viewing evaluation as a continuing system for learning and one part of many systems that provide information and learning opportunities for organizations.

Individual, important evaluation studies will continue to take place. But evaluators have moved from a comparatively narrow focus on methodological issues in the early years to today's broader consideration of the role of evaluation in organizations. Evaluators have recognized that they need to know more about organizational culture, learning, and change, drawing from other disciplines in addition to their knowledge of evaluation theories and practices. They need to identify ways to create an openness to evaluative information and to improving organizational performance, not just the performance of an individual program or policy. As evaluators think of organizational change and learning, they become involved in evaluation-related activities such as planning, performance monitoring, and even fiscal and budgetary decisions. They recognize the need for cooperation across departments or systems that address these related issues so that those gathering and providing information are not working at cross-purposes, but, instead, are collaborating and learning from each other about the information they collect and the methods they use to disseminate information and get it used. Preskill and Boyle (2008) write about the need for organizations to develop "an integrated knowledge-management system" (p. 455) that is aligned with other information systems in the organization. Such systems are essential for many reasons, but reflect the need for planning across systems to maintain information for learning and decisions in the future.

The role of evaluation vis-à-vis schools, organizations, government agencies, and funding sources is changing and will continue to change due to the trends we have discussed here. Evaluation is expanding and becoming more important in the twenty-first century as the world faces critical economic and social challenges. Policymakers, managers, and the public now expect and demand evaluative information, though they may call it by different names. As more people become involved in evaluation within organizations, evaluators will play a critical role in

helping plan systems, build internal capacity, and use methods and approaches that will allow evaluation, or the collection of information to inform and make judgments, to achieve organizational learning.

Table 2.1 summarizes some of the historical trends we have discussed here.

TABLE 2.1 ***Stages in the Development of Evaluation***

Period	*Studies/References*	*Characteristics*
Pre-1800	Sailors eating limes	Most judgments based on religious, political beliefs
1800–1940	Commissions Mass. reports on schools Thorndike and Tyler in ed. Taylor and efficiency Accreditation (Flexner)	Measurement and use of experts begins Focus on public health, education Formal testing begins in schools Social scientists move to government Studies explore social science issues
1940–1963	WW II research on military National Defense Ed. Act (NDEA) Cronbach (1963)	Social science research methods increase Evaluations in schools increase to compete with the Soviet Union Evaluation expands to many areas Methods continue to rely on social science
1964–1973	ESEA of 1965 Head Start Evaluation Great Society Programs Campbell and Stanley (1966) Stufflebeam and CIPP (1971) Stake and Responsive Evaluation (1967)	First mandates for evaluation with Great Society programs A period of social experimentation Texts and articles in evaluation emerge Theorists develop first models Graduate programs in evaluation begin
1974–1989	Joint Committee Standards (1981) Utilization-Focused Evaluation (Patton, 1978) Naturalistic Evaluation (Guba and Lincoln, 1981)	Professional associations, standards, and ethical codes developed Federal support for evaluation declines Evaluation approaches and settings diversify
1990–present	Empowerment Evaluation (Fetterman, 1994) AEA Guiding Principles (1995) United Way Outcomes Measurement System (1996) Participatory models (Cousins and Whitmore, 1998) Third Edition—Joint Committee Standards (2010)	Evaluation spreads around the globe Participative and transformative approaches Theory-based evaluation Ethical issues Technological advances New people conducting evaluation Outcomes and performance monitoring Organizational learning

Major Concepts and Theories

1. Commissions to report on specific problems, objective tests, and accreditations were among the early forms of evaluation. During the Depression, social scientists began working for the federal government to advise it on ways to cure social ills and improve the economy.

2. The Russians' launch of *Sputnik I* created unease in the United States about the effectiveness of techniques used to teach math and science to American students. Congress passed the National Defense Education Act (NDEA) of 1958, which began much evaluation in the educational arena.

3. During the 1960s and 1970s, with the Great Society legislation of the Johnson administration, the federal government began mandating evaluation in many education and social settings. This era of social experimentation represented the first major phase in the growth of evaluation in the United States.

4. The growth of evaluation spurred the first efforts to train and educate professionals specifically to conduct evaluations. Different evaluation theories, models, and concepts to characterize and guide evaluation work began to emerge.

5. The profession became more fully established with the creation of professional associations such as the American Evaluation Association, standards for evaluation, and codes of conduct.

6. The field expanded its methods to include more qualitative approaches and discussions of how evaluators can ensure that evaluation is used by many diverse groups.

7. Since 1990, several trends have influenced evaluation, including its spread to many different countries, more managers and professionals within the organization performing evaluation tasks, a focus on measuring outcomes, and consideration of ways evaluation can influence organizational learning. The tasks for evaluation begin to merge with other areas, including performance monitoring and planning.

Discussion Questions

1. How did the early years of evaluation, before 1965, affect how we think about and practice evaluation today?

2. The Elementary and Secondary Education Act of 1965 (ESEA) and many Great Society programs required that agencies receiving funding submit evaluation reports documenting program results. Discuss the effect of requiring evaluation reports, the impact this mandate had on modern program evaluation, and the problems with both evaluations and evaluators this mandate brought to the surface. What were some important characteristics of evaluation during this period?

3. Since the 1990s, many managers and professionals within organizations have assumed performance monitoring and evaluation responsibilities. What are the strengths and weaknesses of this change? Contrast the knowledge and skill these people bring to evaluation with those of the social scientists who performed many of the mandated evaluations in the 1960s and 1970s.

4. Which of the recent trends we described do you think will have the most impact on evaluation in the future? Why?

Application Exercises

1. What do you see as the critical events and themes in the history of evaluation? How did they shape how people in your field view evaluation? How do people in your field approach an evaluation study?
2. Read one of the interviews cited in the "Suggested Readings" and discuss how this person's experience in the early years of evaluation influenced the field today. How did this influence how you think about evaluation?
3. How has performance measurement or standards-based education influenced work in your school or organization? Are these evaluation measures useful for your organization? For consumers? Why or why not?
4. How does the culture of your organization support organizational learning? How does it support evaluation?
5. Does your organization measure outcomes? Was the focus on outcomes prompted by a mandate, or did your organization choose this focus? How has examining outcomes affected your organization? Its learning?

Suggested Readings

Madaus, G. F., & Stufflebeam, D. L. (2000). Program evaluation: A historical overview. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation.* Boston: Kluwer-Nijhoff.

Mark, M. (Ed.). (2002). *American Journal of Evaluation, 22*(3). This issue contains 23 articles by leaders in the evaluation field on the past, present, and future of evaluation. It is a follow-up to the 1994 issue of *Evaluation Practice, 15*(3), edited by M. Smith, in which different contributors considered the past, present, and future of evaluation.

In 2003, the Oral History Project Team, consisting of Jean King, Melvin Mark, and Robin Miller, began conducting interviews with people who were in the field in the United States in the early years. These interviews were intended to "capture the professional evolution of those who have contributed to the way evaluation in the United States is understood and practiced today" (2006, p. 475). They make for interesting and exciting reading in conveying the nature of evaluation in its early years and its impact on the practice of evaluation today. The interviews are listed in the column to the right. We encourage you to read some of them to gain some insight.

Datta, L. E., & Miller, R. (2004). The oral history of evaluation Part II: The professional development of Lois-Ellin Datta. *American Journal of Evaluation, 25*, 243–253.

Patton, M. Q., King, J., & Greenseid, L. (2007). The oral history of evaluation Part V: An interview with Michael Quinn. *American Journal of Evaluation, 28,* 102–114.

Sanders, J., & Miller, R. (2010). The oral history of evaluation. An interview with James R. Sanders. *American Journal of Evaluation, 31*(1), 118–130.

Scriven, M., Miller, R., & Davidson, J. (2005). The oral history of evaluation Part III: The professional evolution of Michael Scriven. *American Journal of Evaluation, 26,* 378–388.

Shadish, W., & Miller, R. (2003). The oral history of evaluation Part I: Reflections on the chance to work with great people: An interview with William Shadish. *American Journal of Evaluation, 24*(2), 261–272.

Stufflebeam, D. L., Miller, R., & Schroeter, D. (2008). The oral history of evaluation: The professional development of Daniel L. Stufflebeam. *American Journal of Evaluation, 29,* 555–571.

Weiss, C. H., & Mark, M. M. (2006). The oral history of evaluation Part IV: The professional evolution of Carol Weiss. *American Journal of Evaluation, 27,* 475–483.

3

Political, Interpersonal, and Ethical Issues in Evaluation

Orienting Questions

1. Why is evaluation political? What are some of the actions an evaluator can take to work effectively in a political environment?
2. Why are communication skills important in an evaluation?
3. What are some of the key standards by which we judge a good evaluation?
4. What are some of the important ethical obligations of an evaluator?
5. What are some of the sources of bias that can affect an evaluation? How might such biases be minimized?

Before we begin introducing you to the different approaches to evaluation and the technical skills for actually conducting an evaluation, it is important to first discuss some fundamental issues that influence all of evaluation practice. Evaluation is not just a methodological and technical activity. Important as methodological skills are to good evaluation, those skills are often overshadowed by the political, interpersonal, and ethical issues that shape evaluators' work. Many a good evaluation, unimpeachable in all technical details, has failed because of interpersonal insensitivity, poor communication, ethical breaches, or political naïveté. Clients have certain expectations about evaluation. Sometimes these expectations are accurate; sometimes they are not. Evaluators need to listen and observe carefully to learn those perspectives and to understand the political environment in which the evaluation is taking place. Stakeholder groups have different perspectives, different interests, and different concerns about the program

and about the evaluation. Evaluators must be skilled in human relations and communication to work with different groups, to facilitate their communication as appropriate, and to make choices about how the evaluation meets the needs of different groups, all within a political context where different groups are struggling for different resources.

Evaluators cannot afford to content themselves with polishing their tools for collecting, analyzing, and reporting data. They must consider how to deal with pressures to supply immediate data or with the misuse of results. They must consider ways to minimize fears or misunderstandings about evaluation, the means for involving different groups in the evaluation and, then, ways to balance their interests and needs. Evaluators need to think about how evaluation reports will be received by different stakeholders; whether the results of the evaluation will be suppressed, misused, or ignored; and many other interpersonal and political issues. Ignoring these issues is self-defeating, because human, ethical, and political factors pervade every aspect of an evaluation study. It is folly to ignore them, labeling them as mere nuisances that detract evaluators from important methodological tasks. Political, ethical, and human factors are present in every program evaluation, and moving ahead without considering them will lead to a poor evaluation regardless of the technical merits of the study. Recall our discussion of the differences in evaluation and research in Chapter 1. Evaluators are working to make an impact on real people, organizations, and societies. To do so, they must not only collect good data, but they must also see that intended audiences are open to using or being influenced by the data. This can be a challenging task!

In this chapter, we deal with three important, interrelated topics: (1) the political context of evaluation; (2) communication between the evaluator and others involved in the study or the program; and (3) ethical considerations and potential sources of bias in evaluation.

Evaluation and Its Political Context

> Was it mere naïveté that accounted for the initial failure of evaluation researchers to anticipate the complexities of social and political reality? These researchers [evaluators] were mentally prepared by the dominant Newtonian paradigm of social science for a bold exploration of the icy [unchanging] depths of interplanetary space. Instead, they found themselves completely unprepared for the tropical nightmare of a Darwinian jungle: A steaming green Hell, where everything is alive and keenly aware of you, most things are venomous or poisonous or otherwise dangerous, and nothing waits passively to be acted upon by an external force. This complex world is viciously competitive and strategically unpredictable because [evaluation] information is power, and power confers competitive advantage. The Darwinian jungle manipulates and deceives the unwary wanderer into serving myriads of contrary and conflicting ends. The sweltering space suits just had to come off. (Sechrest & Figueredo, 1993, p. 648)

This colorful portrayal of evaluators' first forays into the complex and unpredictable environment in which programs are managed and evaluated underscores a critical point: Evaluators work in a political environment. Evaluation itself is a political act—and the professional evaluator who prefers to eschew "politics" and deal only with technical considerations has made a wrong career choice.

From the beginning of modern-day evaluation, evaluators have written about the political nature of the activity. Suchman (1967), Weiss (1973), and Cronbach and his colleagues (1980) all emphasized the political nature of evaluation, underscoring the fact that evaluation of publicly supported enterprises is inextricably intertwined with public policy formulation and all the political forces involved in that process. However, as Sechrest and Figueredo's description at the beginning of the chapter so vividly indicates, researchers moving into the political arena to conduct evaluations during its time of growth in the United States in the 1970s were unaware of the implications that working in a political environment had for their methodological work. (See also Datta and Miller [2004], and Weiss and Mark [2006] for their descriptions of these early evaluations.)

Today, at least partly because the field has had time to mature and gain more experience in conducting evaluations and to consider the factors that influence their success, evaluators are much more aware that they work in a political environment. Nevertheless, perhaps because the training of evaluators tends to emphasize methodology, evaluators continue to be surprised at the political context of their work and a little unsure of what to do in it (Chelimsky 2008; Leviton 2001). Another explanation for at least U.S. evaluators' naïveté about the political world may rest with the disciplines they studied. A study of members of the American Evaluation Association found that the most common fields of study for U.S. evaluators were education and psychology (American Evaluation Association, 2008). Unlike European evaluators (Toulemonde, 2009), few evaluators in the United States were trained in the fields of political science or economics and, therefore, consideration of politics and the political context may be relatively new to them. Shadish, a leader in evaluation theory and methodology who was trained in psychology, remarks on his coming to understand that politics played an important role in evaluation (Shadish & Miller, 2003). He tells of his surprise years ago that people did not choose to adopt a program that had been proven to be quite successful. The occasion prompted him to read and then write an article on policymaking in the *American Psychologist,* the leading psychology journal (Shadish, 1984). He notes that in preparing the article he read "Politics and Markets" by Charles Lindblom, an esteemed political scientist, along with some other important works in economics and political science, and "all of a sudden I realized that the world didn't work around what was effective. It worked on other matters entirely—on politics and economics" (Shadish & Miller, 2003, p. 270).

In this section, we will discuss the reasons why evaluation is political and the nature of that political environment. Then, we will provide a few suggestions for how evaluators can work effectively in a political world.

How Is Evaluation Political?

The term "politics" has been applied so broadly to so many different phenomena that it has all but lost its meaning. It has come to stand for everything from power plays and machinations within a school or organization to political campaigns or relations among governmental agencies. The *Merriam-Webster Dictionary* reflects these different meanings as it defines politics variously as

- "the art or science concerned with government. . . ."
- "the art or science concerned with guiding or influencing governmental policy"
- "competition between competing interest groups or individuals for power or leadership"
- "the total complex of relations between people living in society" (*Merriam-Webster*, 2009)

So, how is the context of evaluation political? It is political in each of the ways cited in this definitions! Evaluation is most often concerned with governmental programs, whether they are programs funded or operated at the international, national, state, or local level.[1] At the international level, the European Commission, the governing body of the European Union, has mandated cross-country evaluations in Europe and, as a result, has introduced evaluation to many European countries, particularly those in Eastern Europe. In the United States, as we noted in the previous chapter, modern evaluation began through mandates from the federal government during the 1970s, but now is actively conducted at all levels. In addition, evaluation is quite active in state departments of education and in local school districts.

Evaluation is, of course, concerned with "guiding or influencing government policy," the second definition, but perhaps of even more importance, evaluators are working with individuals and groups of stakeholders who are also concerned with guiding or influencing governmental policy. These stakeholders want to influence government policy for many reasons, including helping their constituents and improving government and society. However, one reason for their interest in influencing government policy concerns the third definition: These stakeholders are competing with each other for resources, power, and leadership. Evaluations serve executive and legislative decision makers who make decisions about funding programs; about continuing, expanding, or cutting programs; and about policies that influence those programs. Evaluations

[1]In the United States, many evaluations take place in nonprofit organizations, which, by definition, are nongovernmental organizations. Nevertheless, we will consider these organizations governmental, or political, for the sake of this discussion because in the last few decades, as the U.S. government moved to privatization, many social services that had previously been delivered by government agencies were contracted out to nonprofit organizations. These government contracts are a large part of what prompts nonprofit organizations to conduct evaluations, and their interaction with government agencies places them in a similar political context.

also serve program managers and other stakeholder groups who are competing with other groups for funding, for scarce resources, and for leadership in developing and implementing interventions for solving societal problems. Policymakers and managers, and other stakeholders, are competing for resources, power, and leadership, and evaluation is a powerful tool for them to use in arguing for resources for their group or program. Thus, evaluation is part of the political system and operates within a political context.

Finally, of course, evaluations take place in organizations where complex relationships exist among many groups—in a school among parents, teachers, students, principals, and the central office; in social welfare departments among clients, social workers, managers, and policymakers. Evaluations are political because even the most basic evaluation can upset or change these relationships. The evaluator may include different groups in the decision making about the evaluation, the data collection may prompt stakeholders to reveal beliefs or attitudes they had not considered or had not voiced, and the results often illustrate the multiple ways in which the program is viewed and, of course, its successes and failures. Thus, evaluation work, in itself, is political.

Recall that the very purpose of evaluation is to make a judgment about the merit or worth of a program or policy. In this way, evaluation differs from research. Evaluation is not solely the collection of data using social science research methods. Instead, it involves making a judgment about the quality of the thing being studied. As such, evaluation is highly political. Researchers do not make a judgment; they draw conclusions. Evaluators, however, make a judgment. That judgment may be about a part of a program, as often occurs in formative evaluation, or about the program or policy as a whole to assist in summative decisions. But moving from data to judgment also moves evaluators into the political realm. Further, evaluative judgments often include recommendations for change and such changes are political. These judgments and recommendations have implications for the competition between stakeholder groups and individuals for resources, leadership, and power.

Evaluation in a Political Environment: A Mixed Blessing? For many evaluators, an appealing aspect of evaluation is that it allows them to influence the real world of policy and practice. Researchers are more detached from that world. Research may influence policy or practice, but the researcher has no obligation to make that connection. The evaluator does. Evaluations are judged by their utility, and designing and implementing an evaluation that is likely to be used is one of an evaluator's responsibilities. So, in order to achieve use, evaluators must attend to the political context of the program or policy they are studying.

Many evaluators tend to view politics as a bad thing, but we suggest there is a more enlightened view. Thoughtful evaluators of publicly funded programs view politics as the way laws and program regulations are made, the way individuals and groups influence the government, and the very essence of what enables governments to respond to the needs of those individuals and groups. Indeed, without politics, government programs would be less responsive to public

needs, not more. As Carol Weiss has remarked, "Politics is the method we have as a nation to resolve differences and reach conclusions and decide policy issues. We don't always like the way it turns out, but it's an essential part of our system" (Weiss & Mark, 2006, p. 482).

Furthermore, evaluation serves a central purpose in our political system: accountability. Accountability means that government is accountable to the people, to the public that elects its leaders. Eleanor Chelimsky, the former director of the Program Evaluation and Methodology Division of the U.S. government's General Accounting Organization, now the Government Accountability Office (GAO), which provides evaluative information to Congress, argues that "the ultimate client or user of our work is the public" (2008, p. 403). She views evaluation as central to democracy in making the work of government more transparent so that leaders may be held accountable. It is rare, she notes, for nondemocratic countries to have organizations within government that evaluate its work (Chelimsky, 2006).

Thus, evaluators do work in a political context, and evaluation exists, at least partly, to make government accountable. Of course, this system does not always work perfectly, but it is important for evaluators to recognize the potential for their role within this system. That potential is to provide information to the public and to stakeholder groups, be they policymakers, program managers, or groups lobbying for or against a particular cause. However, to do so well requires the evaluator to have some understanding of that system and the complexities involved in evaluators' interaction with it.

One reason why evaluators are sometimes reluctant to become involved in politics is their concern that the strength of evaluative findings lies in those findings being seen as independent or objective. In other words, policy actors and the public value evaluations because they think evaluators and their evaluations are not political, but instead are neutral and, as such, are providing information that is "untainted" by political views and beliefs. (Most evaluators recognize that, in fact, data and evaluations are inevitably influenced by values and that it is impossible to totally remove bias from data collection. We will discuss that issue further in later chapters on data collection. Suffice it to say here that evaluation is often valued by stakeholders because they perceive it to be objective.) Therefore, evaluators can be legitimately concerned with how their work within this political context may affect the perceived objectivity of their work. How can evaluators interact with those in the political environment to make sure their study addresses important issues and that the results get to the right people or groups without harming the perceived independence or objectivity of their work? There is no easy answer to this question, but we will describe several potential ways in which evaluators can work within the political system.

Interacting with the Political System. Vestman and Conners (2006) describe three different positions in which evaluators may interact with the political system:

1. *The evaluator as value-neutral.* In this position, the evaluator tries to protect or separate the evaluation from politics in order to maintain its perceived legitimacy and objectivity. Evaluators are rational methodologists who collect data and

provide it to stakeholders. Judgments about quality are then made by the stakeholders. The evaluator works to remain separate and independent and, thus, maintain the objectivity of the evaluation.

2. *The evaluator as value-sensitive.* In this position, the evaluator works to maintain the technical aspects of the evaluation, the provision of information, as separate from politics. However, the evaluator recognizes that other elements of the evaluation—in particular providing judgments, considering ethical issues, and encouraging democratic values—require the evaluator to learn of and become involved in the political environment.

3. *The evaluator as value-critical.* Along a continuum of whether it is possible and whether it is desirable to separate evaluation from politics, the evaluator taking this position believes that values are inextricably part of politics and that it is critical for the evaluator to become involved in politics to actively articulate those values. Evaluation and politics are viewed from a larger perspective in this third position. Vestman and Conners note that the value-critical evaluator "views politics as something integrated in our everyday life," and so "there can be no separation between evaluation and politics and therefore no neutral value or operational position taken by the evaluator" (2006, p. 235). The evaluator, then, takes an active role in considering what is in the public good, and serves "as a cooperative and structuring force in our understanding of society" (2006, p. 236). (See also Dahler-Larsen [2003])

Most evaluators today recognize that the first position is unrealistic. This position is one that is frequently taken by applied researchers who move into evaluation and are less familiar with the purposes of evaluation and its goals; in particular, the importance of use. Weiss (1998a, 1998b), Datta (1999), and Patton (1988, 2008a) have all noted that a principal reason that the evaluations of the 1970s were not used was the failure of evaluators to consider the political context. Most evaluators today recognize the need to balance the technical aspects of their study with a need to learn more about the political context to see that their evaluation is useful to at least some stakeholders in the political environment and to ensure that the evaluation is one that furthers democratic values of participation and equality. Our view is that the third position has elements of validity—politics and evaluation, at least informal evaluation, are part of everyday life, data collection is not a truly neutral activity, and evaluators should consider the public good. However, we do think it is important to attend to the view that formal evaluation and evaluators conducting those evaluations provide a different kind of information, one that addresses people's concerns with accountability in today's society. It is important that the results of evaluation studies be trusted; hence, the evaluator must pay attention to preserving the validity of the study and the perceived independence of the results. (You will read more on the ethical codes of evaluation later in this chapter. See also Chapter 8 on participatory and transformative approaches as means for achieving these goals.) However, the three positions developed by Vestman and Conner illustrate the types of relationships that can exist between evaluation and politics and the important issues to consider in those relationships. They help us, then, to reflect

on the role one should play in specific evaluations. That role may differ from one evaluation to the next, depending on the context of the evaluation.

In this section, we have tried to make you aware that evaluation does occur in a political environment, why that is so, and how knowledge of the political context can be an advantage to an evaluator in bringing about use and, ultimately, helping improve government and society. We have also illustrated some of the roles or positions that evaluators can take within that environment and the risks and potential benefits of these different roles. In the next section, we will explore in more depth some of the actions that an evaluator can take to work effectively in a political environment.

Suggestions for Working Within the Political Environment

Chelimsky (2008) has spoken of "the clash of cultures" between evaluation and politics. As we have noted, that "clash" often occurs because our training focuses on methodology and those methodological skills are frequently gained in research courses with a positivist focus on assumptions of independence and neutrality. We think of ourselves primarily as researchers and are concerned that, in working in the political environment, we will lose our independence and neutrality. In addition, students in evaluation typically do not receive much training in working with stakeholders or in a political environment (Dewey, Montrosse, Schroter, Sullins, & Mattox, 2008).

The Joint Committee on Standards for Educational Evaluation, however, has long recognized the need for evaluators to attend to the political context. As we discussed in Chapter 2, the Joint Committee on Standards for Educational Evaluation is a coalition, currently composed of 18 different professional organizations in education and psychology that have interests in evaluation. In 1981, the Joint Committee published standards for evaluators and consumers of evaluation to use to judge the quality of evaluations. In 1994, the Joint Committee wrote this standard:

> *Political Validity.* The evaluation should be planned and conducted with anticipation of the different positions of various interest groups, so that their cooperation may be obtained, and so that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted. (p. 71)[2]

Note that the wording of this standard reflects two concerns: (a) learning about the political context, that is, the positions of various interest groups, so that the study may be conducted feasibly and effectively; and (b) avoiding possible bias of the evaluation during the study and misuse after the study is completed.

[2]The 2010 version of the *Standards* broadened this standard to "Contextual Validity: Evaluations should recognize, monitor, and balance the cultural and political interests and needs of individuals and groups" (Joint Committee, 2010). We approve of the new standard and its broader attention to many elements of context, but use the 1994 version here to illustrate particular elements of the political context.

These two concerns suggest the merits of learning more about the political context and the risks of not doing so. A good evaluator learns about the political context, which includes the positions of various interest groups in regard to the program. More broadly, the evaluator takes the time to learn the identity of the various groups who are interested in the program, who have some power or control over it, or who may be opposed to the program for whatever reason. The evaluator learns the perspectives of each of these groups, in regard both to the program and to related issues. What is their history? What are their values? What has formed their interest in or opposition to the program? What interest might they have in the evaluation? How might they use the evaluation and its results in the future? This period of exploration helps acquaint evaluators with the political context of the program in a positive way and provides an important foundation for the future. They become aware of how the evaluation, and the questions it is addressing, might be used by various stakeholders. They can then consider how those stakeholders might be incorporated, for example, into the evaluation process or the dissemination phase.

The standard also conveys the risks the evaluator faces because evaluations occur in a political context. That is, individuals or groups may act to bias the evaluation. It should be no surprise that individuals or groups who are competing for resources or leadership or power should look to the evaluation as a possible threat or, conversely, as a tool they can use to achieve their goals. Of course, in this time when policymakers place a major premium on accountability and demonstrations that desired outcomes are being achieved, managers of programs, school principals, agency directors, and the like want the evaluation to look good, to show that their program is successful. Conversely, there are others, often less readily identifiable, who may want the evaluation to make the program look bad or to suggest serious problems in implementation or in achieving outcomes. So, of course, the evaluator is subject to political pressure. That pressure can take many forms: working to see that the evaluation addresses the outcomes or questions that a person or group desires; suggesting that certain people be interviewed and others avoided or data be collected in ways that they think will provide desired results; manipulating the interpretation or reporting of results for desired ends, be they positive or negative; and, finally, misusing the results, misquoting them, citing "evidence" without context, or purposely distorting findings. Evaluators are pressured by stakeholders in all these ways—and more. Therefore, it is imperative that the evaluator both know the political environment and stakeholder groups and be willing to take courageous positions to maintain the validity or accuracy of the study and the dissemination of results in an accurate manner. We will address some of these issues in our discussion of ethical standards and codes, or expected ethical behavior in evaluation, later in this chapter. Here, we will discuss some of the steps that the evaluator can take to understand the political context and to avoid problems of bias or misuse because of the political milieu.

Eleanor Chelimsky makes several recommendations for reducing the "clash of cultures" and "improving the 'fit' between evaluative independence and the political requirements of a democratic society" (2008, p. 400). She notes that unwanted political influence can occur at any time during the evaluation: during the design

phase; as the study is being conducted; and as results are being interpreted, reported, and disseminated. Her recommendations address each of these phases. Following are her recommendations and, in the narrative following each, a description of how the evaluator might make use of her recommendations to improve the evaluation.

1. *Expand the design phase.* Take the time to learn the political context as described previously. Learn the history of the program and the values and interests of stakeholders who have supported or opposed it. Consider the reason(s) the evaluation has been commissioned, the questions it is intended to address, and how the evaluation and its focus may complement or subvert the interests of stakeholders.

2. *Include public groups in evaluations, when relevant.* As we will discuss in later chapters, evaluations today are typically at least somewhat participatory and, occasionally, involve many different stakeholder groups. The evaluator might include different groups in an advisory or planning task force created for the evaluation, might collect data from the different groups through interviews or surveys, or might include each group in other ways. Gaining the input of different stakeholders can help the evaluation in many ways. It increases the validity and the credibility of the evaluation because the evaluation results present many different perspectives on the program. Seeking the participation of a number of groups or individuals may gain the support, or at least the understanding, of different groups for the evaluation. Finally, the involvement increases the likelihood that the public and related stakeholder groups are aware of the evaluation and its results. This both helps fulfill the accountability function and can help to achieve other types of use. How can an evaluation make a difference if many of those who are concerned with the program don't know of the evaluation?

3. *Lean heavily on negotiation.* Chelimsky makes two important, but contrasting, points on this issue: (a) Talk, talk, talk with others. Many issues can be negotiated if we only continue to talk with the concerned groups or individuals and find room for compromise and change. (Think of Congress. Why does it take so long to pass a controversial bill? Because the legislators are attempting to represent the viewpoints or needs of their constituents, which often differ dramatically across districts.) (b) If the issue in dispute is something that cannot be compromised or threatens the propriety of the evaluation, such as revealing anonymous sources or altering data or results, the evaluator should show "an unwillingness to be intimidated, even when it's clear the outcome may not be a happy one" (Chelimsky, 2008, p. 411).

4. *Never stop thinking about credibility.* The evaluator's strength lies in the integrity of the study, the use of appropriate methods, honest and balanced interpretation of results, and judgments and recommendations based on those results. Evaluators and evaluation units within an organization or evaluation companies that contract for evaluation gain reputations. Since evaluators are working in a political environment, it is important for clients to believe that the evaluations they conduct or that are conducted by their organization or department are credible, even though the results may not always match some stakeholders' or key clients' wishes.

5. *Develop a dissemination strategy.* Chelimsky (2008) strongly believes in evaluation as an important tool for a democratic society, as the means for making the government accountable to the people. Therefore, the evaluator should communicate the results of the evaluation in ways that can be understood by each audience and develop appropriate methods for those results to be disseminated. In local evaluations, results are often disseminated most effectively in meetings—with parents at school PTA meetings, with program staff at staff meetings. Clients and the larger public may be reached through short pieces in newsletters or on web sites. (See Chapter 16 for recommendations on reporting results.)

Let us add a few other suggestions, building on our previous discussion and Chelimsky's recommendations:

1. *Build in time during the planning stage to learn about the political context.* What does your primary client hope to accomplish with the evaluation? Who funded the evaluation and what do they hope to learn? To gain from it? What other individuals or groups have potential interests in the evaluation? (This would certainly include those served by the program and those delivering the program, agencies that fund or set policies for the program, competing or potentially competing programs, and other programs or organizations that serve the same clients.) Take time at the beginning to interview individuals or representatives of these groups and learn their perspective on the program, their concerns or interests in the evaluation, and so forth. (See Fitzpatrick [1989] for a description of an evaluation in which she analyzes the political environment and identifies viewpoints of different stakeholder groups.)

2. *During the planning stage, make sure your client knows that most evaluations find some successes and some failures.* Many clients assume that their program achieves all of its goals and that the evaluation will demonstrate this. At the early stages, we always find an occasion to mention that few programs achieve all their goals and that it is quite likely that we will find that they are doing very well at some things and not so well at others. Since most of our evaluations have some formative components, we add that we should be able to provide information or suggestions on how to improve their program as well.

3. *Think about the politics of your data collection.* Are there groups, individuals, or data sources that some seem to want you to avoid? If so, why? Pilot test some data collection from this source to get a sense for their perspective or the information they might provide. If the pilot test suggests that their input is useful, use your advisory group and your own reasoning to argue to collect data from this source to add to the validity of the study. Think carefully about any method or component of data collection or design that seems to be especially encouraged or discouraged by particular groups. Why are they taking that perspective? Does the perspective tell you something about their values and the kind of information they need or find credible? Or, are there political reasons—hopes for success or failure—that are influencing their suggestions? Remember that you have been selected to conduct this evaluation at least partly because of your methodological expertise. Use that

expertise to advocate and, ultimately, to select the most appropriate methodological strategies for the evaluation.

4. *Include others, your advisory group and other stakeholders, in your interpretation of results.* It is helpful to get other perspectives, but attend to those perspectives. Is an individual's different interpretation of something a useful insight, one that reflects a genuine alternate perspective? Everyone's perspective, including the evaluator's, is influenced by one's values and experiences. But consider to what extent the perspectives you hear may arise from political concerns and examine the real meaning of those concerns.

5. *Seek input from many on the final report(s) and other products for disseminating results.* The final written report should not be a surprise to key stakeholders. They should have seen and heard the results before receiving the final report. You can make sure they have by including them in a review and meeting with them as data are analyzed to get their reactions and to discuss your interpretations and recommendations. These meetings may be far more useful in achieving understanding and change than the written report. Be clear in how your conclusions and any recommendations emerge from your findings. An evaluation is not about finding fault with individual people; rather, it is concerned with identifying the value and the strengths and weaknesses of programs and policies. So consider your wording. Make suggestions for improvement or action where possible, but do so with care, making sure you can defend your conclusions, your judgments, and your recommendations. (See Chapter 18 for more on reporting findings.)

Fortunately, the *Program Evaluation Standards* and the Guiding Principles developed by the American Evaluation Association also provide evaluators with the means to work with many of these political issues. Many evaluators find it useful to share the Guiding Principles with their client and other stakeholders as they begin their work.

Establishing and Maintaining Good Communications

As this discussion on working in a political context indicates, good evaluation work involves much more than knowing how to collect and analyze data. Our recommendations for working in a political environment often concern communicating with stakeholders. But, interpersonal skills and communication are important enough to merit a separate section here. In this section we want to consider how to develop and maintain good relationships with clients and other stakeholders while conducting an evaluation. After citing some of the dangers to evaluation from working with others in a political environment—"the hidden agendas, cooptation of the evaluator, subversion of the evaluation question, sabotage of the design or the measurement scheme, and misuse of results"—Laura Leviton in her Presidential Address to the American Evaluation Association focused her remarks on the problems evaluators themselves present: "Often, evaluations are blindsided and the product is less than it could be because of our own lack of skill in dealing

with people and organizations" (2001, p. 3). Noting that research shows there are many different types of intelligences and that the strength of evaluators is often in analytical skills, she added:

> I think that sometimes evaluators are absolutely dumbfounded at the negative effects of their words and actions on other people. Yet it should be clear that the ability to communicate well and relate well to others is fundamental to negotiating a better and more useful evaluation question, employing better methods with less resistance, and conveying the results more effectively to clients. In other words, our lack of interpersonal intelligence and "people skills" often make the negative syndromes of evaluation worse than they might otherwise be. Equally bad, our lack of people skills prevents us from optimizing our other talents and skills to produce high quality and useful evaluations. (Leviton, 2001, p. 6)

Now, we think that many evaluators do have interpersonal skills but may simply not think to use them in effective ways in conducting evaluations because they are too focused on methodological issues and their role as social scientists. Just as we discussed the evaluators' obligation to learn about the political context of the program they are evaluating, we want to emphasize that communicating effectively with those involved in the evaluation is critical to the success of the evaluation. As Leviton suggests, evaluators must think about their language, learn about the perspectives of others, and involve them—and learn from them—as the evaluation is conducted.

Here are a few of our suggestions for establishing and maintaining good communications during an evaluation:

1. *In planning the evaluation—writing the proposal or preparing the contract—build in time for communication.* Remember to include time for communication through meetings, meetings, and more meetings! Discuss evaluation plans and results orally with key groups first. Allow for dialogue. Listen to how the different individuals you are meeting with react to the evaluation plans and, later, to the results. Communication with others, of course, should not always be in a group setting. The evaluator should remember to take time to chat with those delivering or managing the program when on site. Learn what their concerns are about the evaluation and about the program itself. What are they worrying about? What pressures are they under? Use interim reports and memos to send information to individuals or groups whom you may not encounter frequently, but follow up by phone or in person to talk with them and get their thoughts. The evaluator needn't be co-opted by seeking to communicate with these groups and to hear their ideas. But hearing their ideas, their perspectives, and their experiences with the program and the evaluation is the only way that the evaluator can break down barriers to evaluation and prepare stakeholders to receive the results and see them as credible and useful.

2. *Prepare clients (those who sponsor the evaluation) and other stakeholders for evaluation.* Develop an "evaluation spirit" by talking with all participants about the purpose and benefits of the evaluation. Resistance to evaluation comes naturally to most

people and not knowing what to expect can only increase such resistance. If stakeholders are new to evaluation, or have had previous bad experiences with it, learn more about their concerns and fears. Ask about their previous experiences with evaluation and what they think this one will do. Let them know your views about evaluation and what it can do. As appropriate, provide stakeholders with information on other evaluations or evaluation approaches or on organizational change and learning to illustrate what evaluation and self-examination can accomplish for an organization. Current literature on continuous improvement and learning organizations can be helpful. (See Chapter 9 for a discussion of using evaluation to improve learning in organizations and how to use the research on learning to improve evaluation practice.)

3. *Invite and nurture outside participation.* In evaluating a school program, for example, remember that parents, school board members, and citizens in the community are all potential stakeholders. Their participation not only strengthens the evaluation but also signals that this is an important project. When evaluating programs in health and human services or in corporate settings, the external stakeholders may be different (e.g., citizen groups, families of those receiving treatment, county commissioners, service providers, corporate board members, consumer advocate groups). Learn who the stakeholders are for the program. Evaluators can play an important role in seeking to empower previously disenfranchised groups by bringing representatives of these groups into the discussion. (See our discussion of participatory and empowerment approaches in Chapter 8.)

4. *Seek input from key individuals or groups on evaluation decisions.* Evaluators should not make important decisions alone. While evaluators may be the most expert in methods of data collection, the client and stakeholders have expertise in the program being evaluated and their experiences with it. Their needs and views must be sought and considered. Foster a spirit of teamwork, negotiation, and compromise. Seek input and consult with others at important points, including determining the purpose of the evaluation and developing the evaluation questions, selecting sources and methods for collecting data, developing measures or looking for existing measures, analyzing and interpreting the data, and, of course, considering the implications of the findings. The evaluator should frequently seek input from others on when to disseminate results (not waiting until all is done) and how to do so. Others will know what is most likely to be heard or read, when individuals or groups are interested in results, and which results would be of most interest to them. Watch out for political agendas, though. Don't make assumptions about what people want to know. Instead, talk with them to find out.

5. *Encourage constructive criticism of the evaluation.* Invite stakeholders to challenge assumptions or weaknesses; encourage divergent perspectives. Model a spirit of fairness and openness when critical feedback is given. By encouraging stakeholders to provide constructive, critical feedback on their work and then responding in an accepting and open manner, evaluators can demonstrate the evaluation spirit they hope to see in stakeholders. (See Fitzpatrick and Donaldson [2002] for a discussion

of Donaldson's use of 360-degree feedback during an evaluation to provide a means for program people to comment on the evaluation, just as the evaluator is commenting on the program.)

Following these recommendations can improve responsiveness to the evaluation—and, hence, its subsequent use—and can enhance the quality of the evaluation product itself.

Maintaining Ethical Standards: Considerations, Issues, and Responsibilities for Evaluators

Given that evaluation occurs in a real-world political context and that to carry out a good evaluation, evaluators must learn about the political context and develop good communications with stakeholders, it should not be surprising that ethical problems can often arise for the evaluator. In our discussion of the political context, we noted the political pressures that can emerge to change the purpose of the evaluation or an evaluation question, to select data sources or designs that are more likely to produce desired results, and, of course, to interpret or report findings in more favorable or desired ways. In addition, as one works to improve communication with clients and stakeholders, closer relationships develop and these relationships can present ethical problems. Therefore, evaluators must be sufficiently sensitive to the potential ethical problems that can occur in evaluation so that they recognize the problems when they occur and have some sense for what to do about them. One step in that direction is gaining knowledge about the profession's expectations for ethical behavior in evaluation.

Let us begin this important section on ethical behavior in evaluation with a real-world example of ethical failures in the evaluation of a different type of product. In 2009, the world was facing what some were calling a "financial meltdown." Home values and the stock market were plummeting. Thousands had to leave their homes because of foreclosures. Unemployment was increasing and predicted to reach 10% in the United States. Countries all over the world were affected by the crisis. Though the factors that contributed to the financial meltdown were many, analysts and elected officials were highly critical of the role of rating agencies in this crisis. Rating agencies such as Moody's Investors Service, Fitch, and Standard & Poor's analyze stocks and bonds and assign credit ratings based on their research. Dating back to the early years of the twentieth century, these agencies began conducting research to judge the quality of companies that issue bonds and, through their ratings, to provide information to investors who made decisions based on these ratings, that is, deciding that a company is safe or unsafe for investment. But changes occurred in recent years that affected the quality of these ratings. For more than 50 years, investors paid these companies for their ratings. As the economy worsened in the 1970s, companies issuing bonds began paying agencies for their own ratings. This established a huge, but relatively unnoticed, conflict of interest.

Rating agencies were now rating the bonds of those who paid them. Of course, some bonds continued to receive low ratings, but many were highly rated. In 2007, 37,000 "structured finance products," the complex financial products that are the source of much of the financial turmoil today, received the highest possible ratings ("A Brief History of Rating Agencies," 2009). Today, many of those ratings have been downgraded, but too late for many investors and citizens who are paying for company bailouts. The first suggestion of problems with the rating systems arose in 2001 when Enron Corporation, which had earlier been highly rated by these agencies, defaulted. In 2007, the heads of these rating agencies were called to testify before the U.S. Congress and were heavily criticized for their failure to identify risky investments and warn investors.

Evaluators, like analysts at Moody's, Standard & Poor's, and Fitch, judge the quality of something and provide information for clients and other stakeholders to make decisions based on those judgments. Although our actions are unlikely to collectively threaten the financial stability of the world economy, many elements of our work are very similar. Clients and others—the public—look to evaluators to provide objective, independent judgments about the quality of programs, products, or policies. We use analytic methods to judge the product and assume that the transparency and validity of those methods will substantiate our findings. But just as employees at the rating agencies must interact with real people at the companies they are judging to conduct their analyses, we, too, interact with clients and stakeholders to learn about the programs or policies we are judging. In many cases, our client is the CEO or a manager of the program we are evaluating. Our results may be used by our client or the manager of the program to seek further funding or to make decisions about funding a program, just as the results of bond raters are used by investors and company managers to make decisions about funding a company. The potential for ethical conflicts—conflicts we may not see—is great. The conflicts lie not simply in methodological choices, but in the relationships that develop when research methods are used in the real world. In this section, we will describe some of the ethical problems that evaluators encounter, discuss ethical codes developed to guide practice, and provide some suggestions of our own to help evaluators consider how to behave ethically.

What Kinds of Ethical Problems Do Evaluators Encounter?

Studies of practicing evaluators reveal the types of ethical challenges that evaluators face. Morris and Cohn (1993) surveyed members of the American Evaluation Association and found that nearly two-thirds of the evaluators had encountered major ethical challenges in their evaluation work. Their analysis of the types of ethical violations that members encountered showed these types of problems:

A. Challenges in the contracting phase:
   - Stakeholder has already decided what the findings "should be" or plans to use the findings in an ethically questionable fashion.

- Stakeholder declares certain research questions off-limits in the evaluation, despite their substantive relevance.
- Legitimate stakeholders are omitted from the planning process.

B. Ethical concerns regarding confidentiality or disclosure agreements:
   - Disputes or uncertainties concerning ownership/distribution of the final report, raw data, etc.
   - Although not pressured by stakeholders to violate confidentiality, the evaluator is concerned that reporting certain findings could represent such a violation.
   - Evaluator is pressured by stakeholder to violate confidentiality.

C. Challenges in presenting findings:
   - Evaluator is pressured by stakeholders to alter presentation of findings.
   - Evaluator is reluctant to present full findings for unspecified reasons.
   - Evaluator has discovered behavior that is illegal, unethical, dangerous, etc.
   - Evaluator is unsure of his or her ability to be objective or fair in presenting findings.

D. Ethical concerns after the report is complete concerning misinterpretation or misuse:
   - Findings are suppressed or ignored by the stakeholder.
   - Unspecified misuse by the stakeholder.
   - Findings are used to punish someone (the evaluator or someone else).
   - Findings are deliberately modified by the stakeholder prior to release.
   - Findings are misinterpreted by the stakeholder (Morris & Cohn, 1993, pp. 630–632).

Morris and Cohn's study remains one of the few to empirically examine the ethical challenges that evaluators face in their work. The most frequent category of problems occurred in preparing results: almost two-thirds of the evaluators reported being pressured by stakeholders to alter results. Morris and Cohn draw several interesting conclusions from their study. First, their content analysis of responses revealed that ethical problems "can, and do, arise in every stage of evaluation" (1993, p. 639). Although respondents reported problems at every stage of the evaluation, the most frequently cited problems occurred at the final stages of the evaluation, in presenting findings. These ethical problems generally arise from pressures from stakeholders, typically the client, concerning the product of the evaluation. In other words, stakeholders are less likely to apply pressure as the study is being carried out than with the final product, the evaluation findings, and the report. In fact, clients presumably value the scientific and objective nature of the work they have hired the evaluator to conduct. But, their concerns emerge with the product itself when the results are surprising or disagreeable. When clients or other stakeholders argue with the evaluator over the interpretation of the results or the presentation of findings, the evaluator may be surprised, having conceptualized his or her role as an independent, objective evaluator. Thus, Morris and Cohn note, the stakeholders' pressures, as seen by the evaluator, "undermine the mission of scientific inquiry, which is to seek the truth and

communicate it" and the evaluators "feel pressured to compromise their role as scientists" (1993, p. 639). These conflicts reveal the "clash of cultures" described by Eleanor Chelimsky and discussed at the beginning of this chapter. That is, stakeholders in this political context are competing for resources, power, and leadership. They see the evaluation findings as a tool that they can use to their benefit in that competition. The evaluation is valued because of its perceived objectivity. But, when the findings clash with their needs in the competition, the political context becomes more important to the stakeholder than the continued objectivity or independence of the evaluator's conclusions.

Faced with such ethical conflicts, the evaluator must take a stand to protect the credibility of this evaluation and future ones. The situation is not an easy one. It is relatively rare for stakeholders to ask evaluators to actually change data. And, to bring about use, good evaluators generally seek the input of clients and other stakeholders on the interpretation of results and presentation of findings in draft reports. So, when the client gives feedback, suggesting changes, evaluators may interpret these "suggestions" differently. (What is the nature of the suggestion? Does it result in a major difference in the interpretation? How strongly does the client ask for or even demand changes?) So the request for changes must be interpreted by the evaluator. Of course, in some cases, the ethical challenge would be quite clear: The client demands that the evaluator change major conclusions concerning the quality of the program. In other cases, the client may be asking for what the client perceives as editing changes, but the evaluator sees as watering down the clarity or strength of the judgments made. How does the evaluator handle this more ambiguous ethical challenge? Dealing with the first situation, in which major conclusions on the quality of the program are demanded, the obvious, ethical challenge requires courage and integrity on the part of the evaluator to maintain the validity of the findings. Dealing with the second ethical challenge certainly may, in the end, require courage and integrity, but may initially require careful thought and reflection concerning the intentions of the client's editing suggestions and the ownership of the report, its wording, and its conclusions. Finally, both situations require the evaluator to recognize that an ethical challenge has occurred.

Although the Morris and Cohn study reveals much of interest concerning the types of ethical conflicts that evaluators actually encounter, they are also concerned that one-third of their sample reported they had not encountered any ethical conflicts in their evaluation work. Their concern, rightly so, is that these evaluators are not just operating in safer environments, but, instead, are not recognizing ethical conflicts or challenges when they arise. As Morris and Cohn conclude, "The subjective notions of ethicality held by many unchallenged group members [those not reporting an ethical challenge] differ in systematic ways from those held by members of the challenged group" (p. 635). Since their study was concerned with describing the ethical problems that evaluators encounter, they were unable to explore the reasons for these different notions of ethical behavior. However, they recommend, and we concur, that the differences illustrate the need for education and training for evaluators to discuss and explore the ethical challenges they may encounter—how to recognize and interpret them, and, ultimately, how to deal with them.

One of the few other studies on ethical behavior among evaluators took a qualitative approach by asking a smaller number of evaluators to discuss how they dealt with ethical issues in their work (Honea, 1992). Honea found that these evaluators seldom discussed ethics or values in their work lives. She identified four factors that seemed to inhibit such discussions. Specifically, her interviewees perceived that:

1. They were being ethical if they were following the model of "objective scientist," and lapses in objectivity were viewed as less an ethical than a methodological concern;
2. Participants in evaluation always behave ethically, so discussion of ethics is unnecessary;
3. Being a member of an evaluation team and engaging in team deliberations prevents unethical behavior from occurring;
4. Neither evaluators nor others involved in the evaluation have the time to confront or discuss ethical issues.

These studies suggest that more attention should be given to ethical issues in educating and training evaluators. In the next sections we discuss the professional codes that can be helpful to evaluators in raising their awareness of ethical obligations and in communicating professional obligations to stakeholders.

Ethical Standards in Evaluation

Since the mid-1970s, the field of evaluation has been active in developing different ethical codes or standards. (See Fitzpatrick [1999] for a discussion of the history of ethical codes in evaluation and a comparison to codes in other disciplines.) Currently, the two most prominent codes for evaluation in the United States are the *Program Evaluation Standards* developed by the Joint Committee on Standards for Educational Evaluation (1981, 1994, 2010) and the Guiding Principles for Evaluators developed by the American Evaluation Association in 1995 and revised in 2003.

These two codes differ in purpose. The *Standards* are designed to assist both evaluators and consumers in judging the quality of a particular evaluation. The Guiding Principles are to provide ethical guidance for evaluators in their everyday practice. The *Standards* focus on the product of the evaluation. The Guiding Principles focus on the behavior of the evaluator. Both, however, inform us as to ethical and appropriate ways for evaluations to be conducted. And, as Sanders (1995) observes, there are no conflicts or inconsistencies between the two documents.

Other countries, too, have been involved in developing ethical codes. The Canadian Evaluation Society (1992) and the Australasian Evaluation Society (Amie [1995]) have each developed ethical codes for evaluators. Many European countries, including Switzerland, Germany, France, and England, have adopted ethical codes or standards. The Swiss and German codes draw on the *Standards* of the Joint Committee, as do the African Evaluation Guidelines (Rouge, 2004). Countries in Asia, South America, and Africa are developing codes either as individual countries or as groups (Stufflebeam, 2004a). This activity reflects the many

and different ethical challenges evaluators face in conducting their work. As Hendricks and Conner (1995) noted when the AEA Guiding Principles were first published, the context of evaluation and the ethical principles that are of primary concern differ across countries. Rouge (2004), for example, discusses the development of evaluation codes for the diverse countries in Africa, and how the codes, although beginning with the Joint Committee's *Standards* as a guide, had to be adapted to the different context of politics and governments in Africa. Specifically, given the many authoritarian governments in Africa, the guidelines include protection for evaluators and special considerations regarding political viability and the disclosure of findings. In these countries where evaluation cultures are new, ethical guidelines can be useful in helping to form those cultures.

What are the ethical obligations of evaluators? We will briefly review the ethical components of the *Program Evaluation Standards* and the Guiding Principles here. The more complete text of both documents is presented in Appendix A.

The Program Evaluation Standards. Before moving into a discussion of the *Standards* themselves, let us briefly describe how the *Standards* were developed. When appointed in 1975, the task of the Joint Committee on Standards for Educational Evaluation was to develop standards for evaluators and other audiences to use to judge the overall quality of an evaluation. Today, 18 academic and professional associations belong to the Joint Committee and oversee the revision and publication of the *Standards*.[3] The *Standards* have been approved by the American National Standards Institute (ANSI) and not only have served as a model for educational evaluations in the United States and Canada, but have also been adapted for use in other countries and in disciplines beyond education, such as housing and community development (Stufflebeam, 2004a). The first *Standards* published in 1981 were designed to address evaluation activities in public schools in the United States. The revision in 1994 expanded their purview to other educational settings, including higher education and training in medicine, law, government, corporations, and other institutions.

The developers of the *Standards* and their revisions make use of an unusual "public standard-setting process" in which evaluators, educators, social scientists, and lay citizens review, field test, comment, and validate the standards (Joint Committee, 1994, p. xvii). Daniel Stufflebeam, who has led the development of the *Standards*, notes that a key step in the early stages in 1975 was the decision to include on the Joint Committee not only professional groups that represent evaluators and applied researchers, but also professional associations that represent school administrators, teachers, counselors, and others who are often clients for educational evaluation

[3]These include the American Evaluation Association (AEA) and the Canadian Evaluation Society (CES), as well as the American Educational Research Association (AERA), the Canadian Society for the Study of Education (CSSE), the American Psychological Association (APA), the National Council on Measurement in Education (NCME), and many associations concerned with school administration and education, including the National Education Association (NEA), the American Association of School Administrators (AASA), and the Council of Chief State School Officers (CCSSO).

(Stufflebeam, 2004a). Inclusion of these groups on the Joint Committee led to some contentious discussions about what constituted a good evaluation. However, these discussions helped produce standards that are useful guides for practicing evaluators in designing evaluations and helping clients and other stakeholders to know what to expect from an evaluation. Standards also play a major role in metaevaluations or judging the final product of an evaluation. (See Chapter 13 for more on metaevaluations.)

The Joint Committee defines an evaluation standard as "[a] principle mutually agreed to by people engaged in the professional practice of evaluation, that, if met, will enhance the quality and fairness of an evaluation" (Joint Committee, 1994, p. 3). As such, the *Standards* are important for the reader to consider before we move into a discussion of how evaluations should be conducted, because the *Standards* communicate what the evaluator, in planning and carrying out an evaluation, should consider. They serve as a guide for the evaluator and a means for the evaluator to discuss and reflect on issues critical to the evaluation with clients and other stakeholders.[4]

The Joint Committee developed 30 standards, which are presented in their entirety in Appendix A. Our attention will be devoted here to the five important attributes of an evaluation under which the 30 standards are organized. The identification of these five attributes was, in itself, a quite significant step for the field of evaluation because it signified the major areas of importance in conducting an evaluation. The four areas are (1) utility, (2) feasibility, (3) propriety, and (4) accuracy. The 2009 revision of the *Standards* added (5) evaluation accountability. Note that prior to the identification of these areas, it was generally assumed that evaluations should be judged based on their validity, or accuracy, because validity is the primary means for judging the quality of research (Stufflebeam, 2004a). The identification of the other areas reminded evaluators and their clients that evaluation also needed to attend to other issues, because it was being conducted in the field and for different purposes than research.

To articulate the meaning of the original four areas, let us draw from the Joint Committee's publication of the *Standards* in 1994.[5] Their introduction to each area addresses the following concepts:

> Utility standards guide evaluations so that they will be informative, timely, and influential. They require evaluators to acquaint themselves with their audiences, define the audiences clearly, ascertain the audiences' information needs, plan evaluations to respond to these needs, and report the relevant information clearly and in a timely fashion. . . .

[4]The Joint Committee notes that not every standard is relevant to every evaluation. They recognize that the context for individual evaluations differs and, therefore, the nature of the evaluation differs. The evaluator and others should consider which of the standards are most relevant for guiding or judging an individual evaluation.

[5]In late 2009, the Joint Committee approved new standards to be published in 2010. We have obtained a prepublication list of the new standards, but the discussion and explanation of these standards are to be published in 2010. Therefore, we present the 2010 standards, but will rely on the previous version for a discussion of the original four categories and their meanings.

> Feasibility standards recognize that evaluations usually are conducted in a natural, as opposed to a laboratory, setting and consume valuable resources. Therefore evaluation designs must be operable in field settings, and evaluations must not consume more resources, materials, personnel, or time than necessary to address the evaluation questions . . .
>
> Propriety standards reflect the fact that evaluations affect many people in a variety of ways. These standards are intended to facilitate protection of the rights of individuals affected by an evaluation. They promote sensitivity to and warn against unlawful, unscrupulous, unethical, and inept actions by those who conduct evaluations. . . .
>
> Accuracy standards determine whether an evaluation has produced sound information. The evaluation of a program must be comprehensive; that is, the evaluators should have considered as many of the program's identifiable features as practical and should have gathered data on those particular features judged important for assessing the program's worth or merit. Moreover, the information must be technically adequate, and the judgments rendered must be linked logically to the data. (Joint Committee, 1994, pp. 5–6)

The identification of these four areas of concern reminds us that evaluation is conducted in the field with the intention of providing sound information to others. The first area emphasizes the importance of use to evaluation and identifies some of the steps the evaluator can take to maximize the likelihood that the evaluation will be used. The identification of feasibility as an area of concern reflects the special considerations that must be made because evaluation takes place in real-world settings with real clients and stakeholders. Procedures must be practical and cost-effective. In addition, for the evaluation to be feasible, the evaluator must consider the context in which the evaluation is conducted—the political and cultural interests. Accuracy standards reflect concerns with the scope of the study and the means by which data are collected. The means for addressing each of these three areas will be discussed further in subsequent chapters. Utility standards and use are the focus of Chapter 17, in which we discuss research and theories on the use of evaluation and recommend ways to increase use. Feasibility is addressed in Chapter 14, in which we discuss planning and managing the study. Finally, accuracy is examined in Chapters 15 and 16 where we discuss methodological concerns.

Here we will focus on the propriety area because our primary concern in this chapter is with ethical conduct in evaluation. The specific standards listed under propriety in the new 2010 *Standards* are as follows:

- "P1 *Responsive and Inclusive Orientation.* Evaluations should be responsive to stakeholders and their communities." This standard, as do many in the 2010 edition, emphasizes the evaluator's obligation to be responsive to stakeholders and to consider the many different groups who may have interests in the evaluation.

- "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." External evaluations generally include

a formal agreement, but internal evaluations often do not. The Joint Committee encourages evaluators to develop a formal agreement at the planning stage of each evaluation and to use it as a guide. The guidelines to this standard provide a useful list of the types of information that might be included in a formal agreement.

- "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." The rights of human subjects are understood to include issues such as obtaining informed consent, maintaining rights to privacy, and assuring confidentiality for those from whom data are collected. (See later section on Institutional Review Boards or IRBs in this chapter.)

- "P4 *Clarity and Fairness*. Evaluations should be understandable and fair in addressing stakeholder needs and purposes." New to the 2010 edition of the *Standards* is an emphasis on clarity, recognizing that many different audiences and stakeholder groups have interests in the evaluation and must receive results in ways that are understandable and comprehensible to them.

- "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." Government in the early twenty-first century has emphasized transparency and the wording of this 2010 standard reflects that emphasis, although previous standards have also emphasized disclosing findings to all who are affected or interested within legal boundaries.

- "P6 *Conflicts of Interest.* Evaluations should openly and honestly identify and address real or perceived conflicts of interest that may compromise the evaluation." Conflicts of interest cannot always be totally eliminated, but if evaluators consider potential conflicts of interest and make their values and biases explicit in as open and honest a way as possible, in the spirit of "let the buyer beware," clients can at least be alert to biases that may unwittingly creep into the work of even the most honest evaluators.

- "P7 *Fiscal Responsibility.* Evaluations should account for all expended resources and comply with sound fiscal procedures and processes." This standard has been included in all editions and reflects the important fiscal obligations of evaluations and emphasizes that the proper handling of these fiscal responsibilities, as well as respecting human rights, is part of the propriety of the evaluation.

Note that the *Standards* emphasize quite a few different issues and, thus, illustrate how ethical concerns cross many dimensions of evaluation and should be considered throughout the study. Traditionally, ethical codes in the social sciences focus on the means for collecting data from others; that is, ensuring informed consent, confidentiality, or anonymity, as appropriate, and dealing with other important issues in protecting the rights of individuals when collecting data from them. These standards indicate that ensuring the rights of human subjects is certainly one very important standard in evaluation. But, the propriety standards also

communicate other areas of ethical concern for the evaluator, such as being responsive to many stakeholders; considering the cultural and political values that are important to the evaluation; being clear on agreements and obligations in the evaluation, conflicts of interest, and reports of findings and conclusions; and managing fiscal resources appropriately. The standard on formal agreements attests to the fact that evaluations, unlike research, always include other parties and, therefore, misunderstandings can arise. Typically, an evaluation involves a partnership between the evaluator and the client. Putting agreements in writing and following them, or formally modifying them as changes are needed, provides the evaluator and the client with a means for clarifying these expectations. At the beginning of the process, the evaluator and client can begin by talking about their understandings and expectations and putting them in writing. This agreement then provides a document to use to monitor these understandings about the evaluation and, thus, can prevent the violation of other propriety standards. Clients, for example, may not be aware of propriety issues such as informed consent or the obligation to disseminate results to others. Formal agreements can work to clarify these concerns. The 2010 *Standards* emphasis on clarity and transparency further highlights the fact that evaluation occurs in the public arena where democracy requires attention to many different stakeholders.

Take a minute now to read the complete text of all of the *Standards* in Appendix A to become acquainted with the meaning and intent of each.

The Guiding Principles. The American Evaluation Association's (AEA) Guiding Principles for Evaluators are elaborations of five basic, broad principles (numbered A–E here to reflect their enumeration in the original document):

A. *Systematic Inquiry:* Evaluators conduct systematic, data-based inquiries.
B. *Competence:* Evaluators provide competent performance to stakeholders.
C. *Integrity/Honesty:* Evaluators display honesty and integrity in their own behavior and attempt to ensure the honesty and integrity of the entire evaluation process.
D. *Respect for People:* Evaluators respect the security, dignity, and self-worth of respondents, program participants, clients, and other stakeholders.
E. *Responsibilities for General and Public Welfare:* Evaluators articulate and take into account the diversity of general and public interests and values that may be related to the evaluation (American Evaluation Association, 2004, The Principles section). (See Appendix A for a more complete presentation of the Guiding Principles.)

Systematic inquiry emphasizes the distinction between formal program evaluation and the evaluations conducted in everyday life. Program evaluators, this principle asserts, use specific, technical methods to complete their evaluations. Because no method is infallible, the principle encourages evaluators to share the strengths and weaknesses of the methods and approach with clients and others to permit an accurate interpretation of the work.

The Competence principle makes evaluators aware of the need to practice within their area of expertise and to "continually seek to maintain and improve their competencies, in order to provide the highest level of performance" (American Evaluation Association, 2004, Section B.4). An emphasis on maintaining professional knowledge is a principle common to many professions' ethical codes, serving to remind their practitioners that their education is ongoing and that they have an obligation to the profession to produce work that maintains the standards and reputation of the field (Fitzpatrick, 1999). The 2004 revision of the Guiding Principles specifically addressed the need for evaluators to be culturally competent in the context of the program they are evaluating. Principle B.2 states

> To ensure recognition, accurate interpretation, and respect for diversity, evaluators should ensure that the members of the evaluation team collectively demonstrate cultural competence. Cultural competence would be reflected in evaluators seeking awareness of their own culturally based assumptions, their understanding of the world views of culturally different participants and stakeholders in the evaluation, and the use of appropriate evaluation strategies and skills in working with culturally different groups. Diversity may be in terms of race, ethnicity, gender, religion, socio-economics, or other factors pertinent to the evaluation context. (American Evaluation Association, 2004, Section B.2)

This new principle reflects the recent attention that AEA and professional evaluators have given to the issue of cultural competence, recognizing that evaluators are often responsible for evaluating programs that serve clients or involve other stakeholders who have different cultural experiences and norms than those of the evaluator. To accurately evaluate the program competently, the evaluator needs to consider the context of the program and those it serves. The 2010 revision of the *Standards* also reflects this concern with its emphasis on learning the cultural context. (See the interview with Katrina Bledsoe, in the "Suggested Readings" section at the end of this chapter for her description of an evaluation where the different cultural norms of clients, volunteers, program staff, and managers were critical to evaluating the program and making recommendations for improvement.)

The principle of Integrity/Honesty also mirrors many of the issues articulated in the *Standards*. It addresses ethical concerns regarding negotiations with clients and relevant stakeholders, conflicts of interest, sources of financial support, misrepresentation of findings, and consideration of methods. Let us highlight two issues here: Guiding Principle C.5 explicitly states, "Evaluators should not misrepresent their procedures, data, or findings. Within reasonable limits, they should attempt to prevent or correct misuse of their work by others" (American Evaluation Association, 2004, Section C.5). Further, Principle C.6 notes that, "If evaluators determine that certain procedures or activities seem likely to produce misleading evaluative information or conclusions, they have the responsibility to communicate their concerns and the reasons for them [to the client]" (American Evaluation Association, 2004, Section C.6). These two principles put evaluators in an assertive position to prevent some of the ethical challenges encountered by evaluators in the research by Morris and Cohn (1993) described earlier.

As noted, the *Standards* and Guiding Principles each provide a means for evaluators to convey to clients their professional obligations. The client has hired an evaluator because of his or her autonomy and expertise. Part of that expertise involves the sense of professionalism that comes from knowing and following the ethical standards of the profession. While evaluators have an obligation to inform clients of these *Standards* and Guiding Principles, conforming to them can be in the clients' self-interest as well, by increasing the credibility of the evaluation.

Respect for People corresponds to Standard P.3, "Human Rights and Respect." This principle and the related standard concern expectations about obtaining informed consent from those from whom data are collected and advising participants regarding the scope and limits of confidentiality. The core of this principle is drawn from the ethical codes of many social sciences concerned with collecting data from individuals—for example, the American Psychological Association, the American Anthropological Association, and the American Educational Research Association. New sections of this principle in 2004 focused on the obligation of the evaluator to understand the context of the evaluation, including the political, social, and economic climate of the program and its stakeholders. This addition built on the evaluator's obligation to have cultural competence. However, it also emphasized the understanding that context and its political, social, and economic components were part of showing respect for people, which was the focus of this principle. Principle D also indicated that the evaluator should ensure that those who provide data do so willingly and do not feel forced into participation out of fear that they may lose the services the program delivers if they decline to participate in the evaluation. Respect for people also reminded evaluators of their obligation to be sensitive to ethnic, cultural, and other differences among participants and stakeholders at all stages of the evaluation, from planning the evaluation to reporting its results.

The Guiding Principles represented a change from the standards developed in 1982 by the Evaluation Research Society, an earlier professional association, by including a greater focus on nonmethodological issues (Fitzpatrick, 1999). This is nowhere more evident than in Guiding Principle E concerning Responsibilities for the General and Public Welfare. This principle emphasizes the obligations of evaluators to include "relevant perspectives and interests of the full range of stakeholders," to consider "not only the immediate operations and outcomes of whatever is being evaluated but also its broad assumptions, implications, and potential side effects," to "maintain a balance between client needs and other needs" and to "go beyond analysis of particular stakeholder interests and consider the welfare of society as a whole" (American Evaluation Association, 1995, pp. 25–26). The inclusion of this principle has sparked dialogue about evaluators' obligations to the public. Certainly, no evaluator has a handle on exactly what the public good is, but Principle E reminds us that our obligation is broader than our particular obligation to the client. Practicing evaluators must also consider the needs of society. Our role might be to stimulate dialogue about those needs or to involve stakeholders in considering the implications of program actions. This principle also might prompt the evaluator to call attention to the need to collect data on unintended side effects of a policy or program either on

the direct clients served or on others who may be indirectly affected by the program. Whatever action is taken, Principle E reminds evaluators to attend to the implications of the program for the community and society as a whole.

In fact, Principle E addresses a concern raised by Smith (1983) prior to the emergence of the Guiding Principles and the 1994 *Standards*. He criticized the writing then on evaluation ethics for focusing solely on methodological issues. Smith wrote:

> Much of the work in evaluation ethics (i.e., the moral behavior of an individual as a professional evaluator) which has been done to date has focused on evaluation moral issues such as confidentiality of data, protection of human subjects, proper professional behavior, and so on. Little has been done on program moral issues, such as: Is this mental hospital placing the community at risk by its early release of patients? Is this nursing home meeting residents' physical needs but at the cost of their human rights of privacy, freedom of movement, and individual expression? Is this educational program for talented students enhancing cognitive skills but reinforcing their emotional dependency on special recognition and privileges? (1983, p. 11)

Principle E addresses Smith's concerns by stating that the evaluator does have an obligation to consider the moral or ethical issues that arise as a result of the program itself.

Readers are encouraged to visit the American Evaluation Association web site (http://www.eval.org/Publications/GuidingPrinciples.asp) to download brochures of the Guiding Principles that can be used to acquaint clients and stakeholders with evaluators' professional obligations and to make use of the additional training materials and readings provided there.

Protections to Human Subjects and the Role of Institutional Review Boards

Both the *Standards* and the Guiding Principles emphasize that to behave ethically, evaluators must protect the rights of people from whom they collect data. Institutional Review Boards (IRBs) are committees of five or more peer researchers who review the data collection plans, or protocols, for proposed research and monitor ongoing research to ensure that the rights of human subjects are protected.[6] IRBs are governed by the Office of Human Research Protections (OHRP), which is part of the U.S. Department of Health and Human Services. Since 1991, federal regulations

[6]The words "human subjects" have historically been used in research to refer to the people who are providing data for the study. That is how we use the words "human subjects" here. However, the word "subjects" implies a helplessness or passivity that many find inappropriate in today's research and evaluation endeavors. As do others, we will typically use the word "participants" when referring to the people who provide data for evaluations through completing surveys, participating in focus groups or interviews, permitting observations, etc. We use the words "human subjects" to avoid confusion when citing or discussing work by others who use these words. Thus, for example, IRBs are often called Human Subjects Review Boards.

require that organizations that receive federal funds for research have IRBs to review all research conducted by the organization. The complete set of regulations is available at http://ohrp.osophs.dhhs.gov. The OHRP can suspend federally funded research if institutions are viewed as noncompliant with regulations. Although such suspensions are "extremely rare and highly controversial," the threat prompted institutions to review and tighten many IRB procedures (Oakes, 2002, p. 450).

The guidelines for protecting human subjects emerged from the Belmont Report (1979) where the focus was on biomedical research, but the Belmont Report itself was prompted by congressional and public outrage at the infamous, 40-year Tuskegee Syphilis Study.[7] Other serious violations in the ethics of social science research have also occurred (Humphreys, 1975; Kahn & Mastroianni, 2001; Milgram, 1974). Institutional Review Boards and their regulations were prompted and guided by the recommendations of the Belmont Report concerning protection of human subjects.

For evaluation studies, the common concern of IRBs is determining whether participants have, in fact, given informed consent to participate in a study. Studies using "vulnerable populations," typically children, pregnant women or women who may be pregnant, prisoners, and people with limited capacity are given special attention by IRBs and the regulations, at least partly because they may not be able to give full, informed consent. Many evaluation studies, however, may be exempt from IRB review according to the regulations. Specifically, research in educational settings that is intended to study traditional educational practices is exempted from IRB review, as well as data collected through "educational tests," defined to include surveys, interviews, and observations of public behavior when individuals are not identifiable and data are confidential. However, the individual evaluator should not decide whether his or her study is exempt. Instead, an IRB board should determine if an exempt status is appropriate. This can often be done relatively easily through contact with the IRB or through an expedited review process. In fact, many evaluations are reviewed through an expedited review process that involves the research protocols for the study being reviewed by one IRB member.

In recent years, however, IRBs have drawn some criticism for their stringent review of social science research, with critics arguing that some IRB requirements have jeopardized legitimate research. We acknowledge, from personal experience,

[7]The Tuskegee Syphilis Study or the Tuskegee Experiment was begun in 1932 and continued until it was halted in 1972. The study initially recruited 399 poor African American men who were sharecroppers and had syphilis with the purpose of describing the natural progression of the disease. In the 1940s, penicillin was found to be a cure for syphilis and became a common treatment. This information was withheld from the men in the study and they were left untreated as the study continued for 40 years. The study was halted in 1972 through the efforts of Peter Buxton, a Public Health Service venereal disease investigator. Although his efforts to stop the study began in the late 1960s, he was unable to stop the study through official channels. He went to the press in the 1970s and congressional hearings were held. Many men died during the course of the study. Forty of their wives were infected and 19 of their children were born with congenital syphilis. In 1997, President Clinton made a formal apology on behalf of the U.S. government, which had funded and conducted the study through the Public Health Service. The study prompted the government to create a commission to write regulations for research, which resulted in the Belmont Report.

that individual IRBs do not always provide reasonable feedback and can overstep their boundaries because of lack of knowledge of research and informed consent. Qualitative data collection, where flexibility and adaptation in data collection may be required, can pose particular problems. IRBs may request standardized interview questions for review when the evaluator needs the flexibility to adapt questions to the purposes of the evaluation and the previous statements of the person being interviewed. The National Science Foundation has taken a leadership role in trying to clarify guidelines for qualitative research. Their web site containing Frequently Asked Questions (FAQs) about data collection and ethical reviews is particularly informative for readers with these concerns. (See http://www.nsf.gov/bfa/dias/policy/hsfaqs.jsp#exempt.) Oakes's "Evaluator's Guide to the IRB" provides more details on the history of IRBs and their requirements (Oakes, 2002).

Our concern in this chapter is making sure that data are collected in a way that protects the rights of participants in an evaluation study. It is not only important for evaluators to learn the policies of the IRBs that oversee their work and the federal regulations that govern them, but also to consider voluntarily seeking IRB review. We agree with many who study the ethics surrounding human data collection that it is useful for researchers and evaluators to seek the opinions of others about their data collection. Often, the researcher or evaluator is too close to his or her own study to see something that might be a threat. IRBs can provide useful input from other researchers who are informed on ethical issues concerning data collection from humans.

Confidentiality and Informed Consent. Confidentiality and informed consent are issues that any evaluator collecting data should be aware of and consider when collecting data. Often, confidentiality and anonymity are confused. Anonymity means that no one knows the identity of the person who provided the data. Confidentiality means that the researcher, evaluator, or person developing the data base may have a code that, in other documents, can be linked to a name, but that the identity of people providing the data will not be revealed to others. Obviously, interviews or observations are not anonymous. The person conducting the interview or observation is aware of the identity of the individual. Similarly, when codes are used on surveys to track who has responded and to prompt those who have not responded to do so, someone is able to make a link among those codes, the responses to a survey, and an individual name. However, the data analysis will not make use of the individual identifiers and, hence, the data are confidential. Further, specific procedures for separating the names and codes from the data and for maintaining the security of the list of names and codes must be established. Any data collection activity should correctly inform individuals as to whether the data they provide should be considered anonymous or confidential.

Informed consent is a central mechanism for protecting the rights of human subjects. As Oakes has written in his Evaluator's Guide to the IRB, "Informed consent is one of the primary ethical requirements underpinning research with human subjects" (2002, p. 463). Informed consent emerged as a central ethical

principle after the Nuremberg trials of Nazi scientists who had conducted research on prisoners in concentration camps. The Nuremberg Code, developed after the trials, established the principle that researchers should not collect data from someone without first obtaining their consent and that such consent should be both fully voluntary and informed. "Informed" means that participants should be told of the purpose of the research, its potential risks and benefits to them, the confidentiality of the information and other relevant issues concerning how it will be handled and protected, and what participating in the study will mean for them, that is, the data that will be collected. The voluntary nature of their participation should also be made clear. For example, evaluations are typically conducted in the context of a program. Participants need to know that they can continue receiving the services of the program even if they choose not to participate in the research. If people receiving services believe they must participate in the research to continue in the program, their participation in the research is not truly voluntary. Informed consent is typically obtained through an informed consent form that describes the study, its purpose, the potential risks and benefits, the voluntary nature of participation, how data will be handled, and other relevant concerns. But it is important that the language of such forms be clear and understandable to the intended audience. A member of the evaluation team trained in the ethical concerns of informed consent should be present to answer questions a participant may have. (See Fitzpatrick [2005] on informed consent.) IRBs typically pay considerable attention to the issue of informed consent and may have sample consent forms for new evaluators to use as guides for consent forms in their own research.

Cultural Competence and Sensitivity. Finally, ethical data collection involves sensitivity to the cultural norms and beliefs of the individuals and groups from whom one is collecting data. Information that might be considered quite appropriate to collect in one group might be considered quite private or misunderstood by another. Such sensitivity is part of attaining cultural competence, as illustrated in the Guiding Principle B.2 and also in D.6. Consider, for example, an evaluation of a school program for children of recently arrived immigrants. Some of those immigrants may be in the country illegally; however, in most cases, the immigration status of children's parents would be irrelevant to the evaluation. The evaluator should avoid pressure to collect such data and, more importantly, should consider the wording of individual questions. If the questions appear to be gaining information that pertains to entry to the country, it may threaten the validity of responses to other items and may not show respect for the privacy of those participating in the evaluation. Evaluators should recognize that evaluation and confidentiality may be suspicious or alien concepts to those completing the survey. Providing information on themselves and their families may be a threatening or frightening experience. Needless to say, surveys should be translated into a language and words that parents can read and understand. Interviews should be conducted by individuals who are not only fluent in the language, but have good knowledge of the culture and norms of the particular immigrant group.

In closing, let us emphasize that evaluators have an ethical obligation to consider the rights of those from whom they are collecting data and to make sure that those rights are protected. Seeking input from others—whether it be from an IRB committee or other informed researchers, members of an advisory group, clients, or representatives of the individuals from whom data are collected—should be a central part of that process. Individuals providing data should be informed of the purposes of the research and any risks that may be incurred by participating. Further, the evaluator should collect only the data necessary and essential for the evaluation. An evaluation does not give one license to collect irrelevant data or to unnecessarily intrude on the privacy of individuals.

Learning and Practicing Ethical Behavior. In this section, we have attempted to acquaint the reader with the standards and codes that have been developed to guide evaluators. But applying these standards or codes to individual evaluations is a much different issue. As the Joint Committee emphasizes, not all standards are equally important in every evaluation. Choices must be made. Similarly, while the Guiding Principles are intended to "proactively guide the behaviors of professionals in everyday practice" (American Evaluation Association, 2004, Preface C), deciding how to apply them in a specific evaluation, particularly when conflicts emerge, requires careful consideration and tough choices. Morris and Cooksy have helped make us aware of the complexity of these choices through an ongoing column on ethical dilemmas in the *American Journal of Evaluation.* The column, begun by Morris in 1998 and assumed by Cooksy in 2004, presents an ethical problem and calls upon two different, experienced evaluators to describe how they would respond to the issue. Distinct differences emerge. See, for example, the disagreements between Cooksy and Knott on an ethical problem concerning sexual harassment by a manager that was reported during confidential interviews (Morris, Cooksy, & Knott, 2000). These differences help educate and sensitize evaluators to recognize and analyze ethical problems they encounter and consider the choices they make. We encourage readers to read some of these ethical dilemmas and the responses to consider how they would respond and to refine their skills in ethical reasoning for evaluation practice.

In the next section, we will give more attention to an ethical problem or concern that occurs in every evaluation—bias and its many sources.

Reflecting on Sources of Bias and Conflicts of Interest

Several of the *Program Evaluation Standards* (U.1, U.4, P.6, A.8) and Guiding Principles (C.3 and C.4) are concerned with the importance of evaluations being honest and impartial, avoiding conflicts of interest and conducting evaluations with integrity. Yet, as research has shown, many evaluators do not believe they have encountered ethical problems and think that because they are following accepted social science methods and the model of the "objective scientist" they are, of course, behaving ethically (Honea, 1992; Morris & Cohn, 1993). In this section, we would like to discuss potential biases and conflicts of interest that evaluators must consider carefully.

First, we should acknowledge that the possibility of human beings rendering completely unbiased judgments is very slight. In fact, it is ironic that some evaluators actually could be more susceptible to bias, simply because they believe that by using social science methodologies to draw conclusions, they are objective and unbiased. But, Carol Weiss, one of the founders of evaluation notes, "You never start from scratch. We pick up the ideas that are congenial to our own perspective. Therefore, people pick up this thought or that interpretation of a research report that fits with what they know or what they want to do" (2006, p. 480).

Evaluators and those involved in evaluation should carefully reflect on their biases. By becoming aware of those biases, one can consider and perhaps counteract some of their influence on evaluation. The ethical evaluator recognizes that evaluation practice consists of making choices—choices about evaluation purposes and questions, about which stakeholders to involve and which designs and data collection strategies to use, about ways to analyze data and to interpret the results. Note, for example, that the raters at Moody's or Fitch's who were researching and assigning ratings to bonds were making choices, too—about what information was important and what was not, about what types of business and investment strategies were fiscally sound. Their ratings involved much more than simply adding together some numbers. Swayed by the tenor of the times, changing business practices, the interests of their rating company and those who paid them, their ratings were influenced in undesirable ways. Choices are, by nature, subjective. Evaluators increasingly realize that bias—inadvertent or conscious—can intrude subtly into nearly every choice they make, from selecting an evaluation approach to writing a report. To avoid the faulty findings of bond analysts, evaluators must think more carefully about the potential sources of bias and conflicts of interest that can occur in each evaluation they are conducting.

It is worth noting that when asked to describe ethical problems they have encountered, evaluators tend to describe problems presented by stakeholders (Morris & Cohn, 1993). As Morris and Cohn themselves note, it may be more difficult for evaluators to recognize or report ethical problems that were of their own doing. The only ethical problem they found that appeared to originate with the evaluator was the concern about their ability to be objective or fair in presenting findings. Recognition of this particular problem, though, is a major first step. It suggests that, even when encountering many ethical problems presented by client or stakeholder pressure, some evaluators remain conscious of how their own biases can interfere with the accurate presentation of results. (Of course, stakeholder pressure and concern with being objective and fair in presenting findings can overlap. When faced with strong pressure from a client, it may be difficult not to become biased against that client and overreact and become less fair and objective in the opposite direction. This might result in reporting or emphasizing problems either in retaliation or to show that you are objective, rather than maintaining a balanced view. It can be difficult to see things from the perspective of someone who has behaved inappropriately toward you, yet that is what the evaluator must do in order to consider all sides.)

Guidance from Ethical Codes and Standards. The Joint Committee *Standards* and the AEA Guiding Principles can often serve as a good first step to raising awareness and considering potential problem areas. So, let us review a few of the Standards and Principles that are relevant to the issues of bias and conflict of interest. Guiding Principle C concerns integrity and honesty. Principles C.2 and C.4 directly address expectations concerning values, interests, and relationships:

> **C.2** Before accepting an evaluation assignment, evaluators should disclose any roles or relationships they have that might pose a conflict of interest (or apparent conflict of interest) with their role as an evaluator. If they proceed with the evaluation, the conflict(s) should be clearly articulated in reports of the evaluation results. . . .
>
> **C.4** Evaluators should be explicit about their own, their clients' and other stakeholders' interests and values concerning the conduct and outcomes of an evaluation (American Evaluation Association, 2004, Section C Integrity/Honesty).

Principle C.7 addresses expectations concerning financial disclosures:

> **C.7** Evaluators should disclose all sources of financial support for an evaluation, and the source of the request for the evaluation (American Evaluation Association, 2004, Section C, Integrity/Honesty).

Several standards also reveal critical expectations for the evaluation itself in regard to the nature of reporting information, the credibility of the evaluator, and the identification of the values involved in interpreting the findings of the evaluation and making final judgments. One such standard is the Propriety Standard P.6, on Conflicts of Interest described above. The Joint Committee defines conflict of interest in this way: "Conflict of interest exists in an evaluation when the personal or financial interests of an evaluator might either influence an evaluation or be affected by the evaluation" (Joint Committee, 1994, p. 115). They note that such conflicts can be caused by "close friendships and personal working relationships" that are more common in internal evaluations and by external evaluators' desire to gain future contracts (Joint Committee, 1994, p. 116). We will discuss interpersonal and financial conflicts of interest later, but, first, we will focus on the standards that provide direction to evaluators.

In describing elements necessary for an evaluation to attain the Accuracy Standard, the 1994 edition of the *Standards* specified Impartial Reporting as an important standard:

> • **A.11** *Impartial Reporting.* Reporting procedures should guard against distortion caused by personal feelings and biases of any party to the evaluation, so that evaluation reports fairly reflect the evaluation findings (Joint Committee, 1994, p. 181).

The 2010 revision of the *Standards* continues this emphasis:

- **A.8** *Communication and Reporting.* Evaluation communications should have adequate scope and guard against misconceptions, biases, distortions, and errors (Joint Committee, 2010).

Interestingly, the other two standards that address bias fall under the Utility Standard, reflecting how important transparency and credibility are to the ultimate use of the evaluation study and its results:

- **U.1** *Evaluator Credibility.* Evaluations should be conducted by qualified people who establish and maintain credibility in the evaluation context.
- **U.4** *Explicit Values.* Evaluations should clarify and specify the individual and cultural values underpinning purposes, processes, and judgments (Joint Committee, 2010).

Bias Introduced by Values, Views, and Culture. These principles and standards are helpful in reminding the evaluator and the client and other stakeholders of professional expectations. Although the standards and principles are worded in terms of what is expected, they indirectly attest to the harm that unacknowledged values and relationships—between people or organizations—can do to the credibility and integrity of an evaluation study. Therefore, evaluators should seek to consider not only the values of stakeholders and clients, but also their personal values that influence both the conduct and the outcomes of the evaluation. What are the evaluator's views, values, and experiences concerning the program or others like it, its clients, the organization in which it resides, and its mission? Suppose you are called upon to evaluate the success of a school serving low-income immigrant students at achieving state academic standards. The school has failed to meet the acknowledged high standards for the past 3 years and is now under review. It could be closed in the following year and students could be moved to other schools. What are your views on educational standards? On high stakes testing? On the efforts of this school, its teachers, and its administrators? On the children it serves? How will your values and views affect how you conduct the evaluation? The stakeholders you include in the evaluation? The way in which you interpret the results? Your ultimate conclusions? Will you be able to report results impartially? Standards-based education and its policies to raise achievement are controversial issues in the United States. Almost everyone concerned with education has a point of view and experience with standards. It would be almost impossible to avoid having these views and experiences affect at least some of the ways in which you conduct the study and reach your conclusions. What steps would you take to attempt to reduce the impact of your views and experiences so that the evaluation is not biased and is seen as credible? Or, do you view it as fortuitous that you have been asked to conduct this evaluation, perhaps because past state or local studies on this issue have been conducted primarily by people from "the other side" (whatever that side may be), people unaware of ethical codes in evaluation who have allowed their

views or experience to influence the work? Revealing your views might jeopardize your opportunity to conduct the study and, thus, present different points of view. What should you do?

Another example of the difficult issues one can confront when considering the bias one's own values introduce might be helpful. You have been asked to conduct an evaluation of support groups for children who have had a parent die. You think you might be able to do an especially good job at evaluating such programs because you have personal experience with this issue. Your spouse died unexpectedly when your children were relatively young, and they participated in grief groups for children and teenagers. You have also read quite a bit on the issue and know what helped your children. Will your views and personal experience with grief groups for children enhance your ability to conduct the evaluation or detract from it? Are you obligated to tell the client about your personal experience and views? How much of your personal experience or your children's personal experience are you obligated to reveal?

Cultural competence, or cultural incompetence, is another personal factor that influences the validity and ethicality of an evaluation. Kirkhart has discussed our own difficulty in seeing our "cultural boundedness"; yet, a good evaluation should describe "multiple cultural perspectives accurately, soundly, and appropriately" (1995, p. 3). As noted, the 2004 revision of the Guiding Principles spoke to the importance of cultural competence. We will discuss cultural competence more fully in Chapter 9, but it is essential to address the issue here as well. Cultural competence has emerged as a concern in evaluation because of the recognition of the role of one's own values and experiences on the conduct of an evaluation. Hood notes that "[t]he evaluation community is replete with those who have limited understanding of the values that are grounded in the racial and cultural backgrounds of groups other than their own" (2000, p. 78). Many of the people served by public or nonprofit programs are people in need. They are likely to differ in many ways from the evaluator: obviously in income; possibly in race or ethnicity; perhaps in their goals and the values, beliefs, and expectations they have in regard to the program; and quite probably in how others treat and view them.

Strategies for Reducing Bias. What can an evaluator do to minimize the bias that personal views and experience bring to an evaluation? One strategy recommended by qualitative researchers (Lincoln and Guba, 1985, Miles & Huberman, 1994; Schwandt & Halpern, 1989) is to maintain an "audit trail," which Schwandt defines as "a systematically maintained documentation system" (2001b, p. 9) to record all the details of the process of conducting the study. The audit trail would include the evaluator's notes on evolving perceptions, day-to-day procedures, methodological decisions, day-to-day personal introspections, developing insights and hypotheses to help the evaluator explore how the evaluation design is emerging and the values and experiences that may influence the evaluator in that evolution. (See Cooksy [2000] for an excellent example of using such memos to aid in reflecting on an ethical problem encountered in data collection.) The evaluator may choose to use the notes for self-reflection and consideration of how

values and experiences may be introducing bias. Alternatively, the evaluator may decide to share portions of the notes with an external party. This person, generally another evaluator, can review the audit trail to explore the appropriateness of the evaluation decisions and the ways in which bias may have been introduced.

Another strategy for minimizing bias is through the process of metaevaluation, or the evaluation of an evaluation in which an outside person reviews an evaluation for its quality. This topic is addressed in detail in Chapter 16. Whatever methods are used, it is important for evaluators (and clients) to examine their personal values and beliefs and consider how these factors can influence their approach to each evaluation and to their eventual conclusions and judgments. Becoming aware represents the first step in preventing bias.

Interpersonal Relationships and Bias. It is apparent to even the casual observer that individuals' feelings toward one another can color their judgments, not only about each other but about practically anything with which the other person is perceived to be associated. Hence, we have legal restrictions on testimony about one's spouse and anti-nepotism policies that prohibit individuals from being in positions where they would need to make decisions about the salary, promotion, or job security of a family member. Similarly, evaluators should avoid evaluating programs that a close friend or family member is concerned with, whether as a policymaker, a manager, or a person delivering the program. The apparent conflict of interest would be too strong even if the evaluator were able to overcome the bias the interpersonal relationship introduced.

Internal evaluators, except in the largest organizations, are almost inevitably evaluating programs that are staffed by someone they know. Therefore, internal evaluators need to think carefully about how to define their role in such settings. Even if the purpose of the evaluation is formative or for organizational learning, the evaluator needs to be prepared to give negative feedback. To achieve change, that feedback may be given in a way that is clear but palatable. Nevertheless, evaluators should be alert to examining how their relationships with those who operate or manage the program can influence the choices and decisions made. Such relationships can affect many elements of the evaluation, from the questions the evaluation addresses to the ways in which results are interpreted and presented. As an evaluator, you are hired or assigned to provide an independent, impartial judgment, and concerns about personal relationships should not interfere with the evaluator's responsibility.

As we discussed earlier in this chapter, however, evaluators have a responsibility to develop some type of relationship with the client and stakeholders concerned with the evaluation. They must be able to communicate with them effectively so they can understand their needs and provide information in a way that meets those needs. Evaluators who are entirely new to the setting of the evaluation should spend time observing the program, meeting with clients and stakeholders, and developing relationships. These relationships are intended to help the evaluation to succeed—to reduce mistrust, to improve understanding, and so

forth—but these relationships also introduce bias. Evaluators are likely to feel more comfortable with people whose values and beliefs are like their own, who support the evaluation and who are open to its methods and interested in its results. At the same time, evaluators learn that some people concerned with the evaluation are more difficult. They are suspicious, accusatory, demanding, inflexible, or behave in any number of ways that are frustrating to the evaluator. These relationships, good and bad, influence the evaluator's behavior. Is the evaluator prepared to give tough, negative results—ones the evaluator knows they won't like—to people with whom he has established rapport? To those who are helpful to the continuation of the study? These are tough issues, but evaluators must be prepared to deal with them, to prepare audiences for difficult results, and to prepare themselves for delivering them.

We find it is useful to clarify and demonstrate one's role at the beginning of the evaluation, during the planning phase. Evaluators do not need to be the tough guy but they do need to be willing to ask tough questions and provide difficult feedback. In a later chapter, we will discuss using logic models during the early stages of the evaluation, as a method to help the evaluator understand the program. This is often a useful time to ask probing or tough questions such as, "Now, why is it that you think that this activity will lead to X change? Some of the research I have read doesn't support that," or "Which of your objectives do you think you are probably not achieving?" You may choose other questions, but our point is that at the beginning of the study—not waiting until the end—you should start to define your role as someone who is interested in them and their program, but is also curious, objective, and questioning. This persona or manner then becomes part of your interpersonal relationship with others in the program.

Financial Relationships and Bias. Unfortunately, financial considerations are a source of bias in evaluation just as they were in the rating of bonds by Moody's discussed earlier in this chapter. We doubt there are many instances of evaluators being bribed to sway an evaluation one way or another, but financial pressures are rarely so obvious and direct. To illustrate how thorny this situation can be, let us describe an actual case.

An evaluator of our acquaintance—we'll call her Diane—was employed by a U.S. government-supported research center whose mission was to develop and test exemplary programs and practices for schools. Assigned to direct the center's evaluation unit, in due time Diane completed an evaluation of a center program designed to improve secondary school students' mathematics performance and attitudes toward math (AMP). The AMP program was expensive. Congress had invested more than $1 million in its development and, although Diane found that students liked the program, there wasn't a shred of evidence to suggest that it had any impact on their performance. Troubled by the implications of reporting such information to the funding agency through which Congress had initiated the program, Diane finally worded her draft report to convey that the evaluation was

to blame for AMP's failure to produce evidence of success. The summary of her report read as follows (italics ours).[8]

> *Summary.* The results of this study indicate that the Accelerated Mathematics Program (AMP) was somewhat effective in developing positive attitudes toward mathematics, in the sense that students tended to like the AMP materials. The study supplied no evidence, however, from which either long- or short-term student performance changes in mathematics ability can be inferred. *The results do not necessarily indicate that AMP was not effective in promoting change in math performance, but that a variety of shortcomings and limitations of the evaluation design did not allow for the identification and measurement of these changes.*

And how did the funding agency respond to this obvious effort to soften the bad news? Their reaction to the draft report came in a letter, which is reprinted here.

> Dear Diane:
>
> Thank you for the three draft copies of the AMP Impact study. I look forward to the final report.
>
> I hope that our future efforts will be structured so that statements such as those in the "summary" will not have to be made. Instead, I hope that we will be able to say something positive in the final report about changes in important performances. I have heard so many good things about AMP that I am disheartened by the lack of evidence that it has short-term performance effectiveness and that I cannot therefore argue for its potential for long-term effectiveness.
>
> The issue here is straightforward. The best argument for funding centers such as yours that I can make internally here in the Department and externally with the Congress is that our products lead to measurable changes for good in American schools. Regardless of the positive "feelings" I get about AMP, it appears we cannot justify all the effort in terms of performance criteria, as per your draft report. That is a drawback, but one which I think we can overcome in future efforts, hopefully in your final report.
>
> Sincerely,
>
> *Larry*
>
> Lawrence T. Donaldson
> Chief Administrator

The message is blatantly clear. Diane better find something positive to prove AMP and its cohort programs are worth the investment, or funding could be withdrawn, the program would fold, and Diane herself would be looking for other employment. It would take a robust soul indeed not to feel some ethical strain in such a situation, especially when her salary comes directly from the threatened program!

[8]The names, organizations, and titles in this summary and the following letter have been changed to provide anonymity, but the essential content has not been altered and is reproduced here verbatim.

Fortunately, though Diane equivocated at first, this story eventually had a happy ending. The final report told the true story, and Diane was able to assume the role of evaluator (with a clear conscience) on the development staff for another program at the same center.

Even when the evaluator is external to the agency whose programs or products are being evaluated, financial dependence can be a potential source of bias. Consider, for example, the delicate balance that must be maintained by external evaluation consultants or firms who are inevitably dependent on repeat business. Scriven (1993) points out this potential source of bias succinctly: ". . . one key economic insight about evaluation contracting is this: No one ever got rich from one evaluation contract" (p. 84). The possibility of future evaluation contracts or consulting depends on how well the client likes the most recent evaluation completed by the evaluator. No problem here if the client has a penchant for the truth, even if it might reflect negatively on the program. But what if the client goes rigid at the first hint of criticism? Developing formal agreements as indicated in the *Standards* can provide some assurance, but evaluators should always think carefully about financial relationships and recognize that their long-term reputation as an independent, impartial evaluator is critical to their sustainability.

Organizational Relationships and Bias. Organizational relationships may be of greater concern to evaluators than immediate financial gain. The relationship between evaluators and the programs they evaluate can determine not only their present financial welfare but their future employment as well. Further, an organization may exert great (or total) control over the evaluator's other perquisites: such things as office space; access to resources, facilities, and record keeping systems; even the convenience of available parking space. The way the organization exercises this control to make the evaluator's life easier or more difficult can certainly cause problems with bias.

To make this point, we present in Table 3.1 eight possible organizational relationships between evaluators and the program being evaluated. Generally, the greatest potential of bias exists in the first row of Table 3.1, and the least potential of bias exists in the last row. Thus, the potential for organizational pressure is greater when the evaluator is employed by the organization whose program is being evaluated than when the evaluator is employed by an outside agency. In addition, bias is more likely when the internal evaluator reports to the director of the program being evaluated than when the evaluator reports to someone outside that program. Sonnichsen (1999), the director of the internal evaluation unit at the FBI, argues that internal evaluators must be placed independently, separated from programs, to be effective. Lovell (1995), in commenting on internal evaluation, notes that, in the long term, the organization expects internal evaluation to pay off, that is, to provide recommendations for improved organizational operations. Bias that produces overly positive reports on programs leads to evaluation not fulfilling its promise.

Mathison (1999) has served as an internal and external evaluator and has written often on the issue. She believes that internal and external evaluators face the same ethical challenges but are part of different communities and that these

TABLE 3.1 *Organizational Relationships of Evaluator to Client*

Evaluator Employed	*To Do One Evaluation or Successive Evaluations*	*Evaluator Reports*
1. Within organization which has responsibility for the program being evaluated	1. Successive evaluations	1. Directly to director of program evaluated
2. Within organization which has responsibility for the program being evaluated	2. One evaluation	2. Directly to director of program being evaluated
3. Within organization which has responsibility for the program being evaluated	3. Successive evaluations	3. To someone outside the program being evaluated but within the same organization
4. Within organization which has responsibility for the program being evaluated	4. One evaluation	4. To someone outside the program being evaluated but within the same organization
5. By outside agency	5. Successive evaluations	5. As consultant or contractor to director of program being evaluated
6. By outside agency	6. One evaluation	6. As consultant or contractor to director of program being evaluated
7. By outside agency	7. Successive evaluations	7. Directly to outside funding agency which supports the program
8. By outside agency	8. One evaluation	8. Directly to outside funding agency which supports the program

communities influence their responses to ethical challenges. Internal evaluators, she asserts, operate within fewer communities than the external evaluator and their primary community is the organization in which they work. Consider, simply, the amount of time the typical internal evaluator spends in that organization over weeks and years. External evaluators, in contrast, have many communities, which can include the organizations they evaluate, the organization that employs them, colleagues in evaluation and their professional association, funding agencies, and others. These communities, Mathison argues, influence evaluators' ethical choices in many complex ways. For example, the internal evaluator's closeness to the organization and relationships in it may enable her to behave more ethically when it comes to creating an ongoing evaluative culture in the organization or sustaining a dialogue about a controversial issue uncovered by an evaluation—a dialogue that may be required to bring about change. In contrast, the external evaluator's diversity of communities and greater distance from the community of the organization being evaluated makes it easier for her to raise questions concerning unethical issues in the program or in the organization. Mathison's concept of the communities of reference for internal and external

evaluators is useful in considering the types of problems each can deal with more effectively, and it helps us to recognize the complex influence of personal, interpersonal, financial, and organizational factors on the ethical behavior of evaluators.

A final important consideration when considering the influence of organizational and financial relationships on bias is whether the evaluation is primarily formative or summative. In considering the pros and cons of an evaluator's financial and administrative dependence or independence from the client, such dependence may be not only tolerable in a formative evaluation, but even desirable. The internal evaluator's relationship with the organization may prompt him to be more responsive to particular information needs of the program and the organization because of his greater understanding and loyalty to the organization. Or, as Mathison notes, the internal evaluator's close relationship with the organization can prompt him to sustain dialogue on an issue long after the external evaluator has gone and to improve the nature of that dialogue because he knows the values and beliefs of those in the organization. However, an internal evaluator may not be so effective for a summative evaluation, particularly if the evaluation concerns large, costly, or high-profile programs. In this case, the internal evaluator's relationships with the organization and its employees, especially if the internal evaluator is affiliated with the unit operating the program, are quite likely to introduce bias. An external, independent evaluator is generally to be preferred in summative evaluations of this type. As we have noted in the prior section, though, independence is defined by a variety of factors.

Ethics Beyond a Code of Ethics

The evaluation standards and guidelines described earlier are, in our judgment, singularly useful in improving the practice of evaluation. We urge anyone aspiring to do high-quality evaluation to become intimately familiar with those standards and guidelines and to apply them diligently. At the same time, mere adherence to ethical standards, however sound, does not ensure ethical behavior. As Peter Dahler-Larsen has written in regard to the broader issue of codifying evaluation practices, these codes serve "at best, as aids to a competent judgment in evaluation, not substitutions for it" (2006, p. 154). Mabry (1999) reminds us that codes of ethics don't remove the subjectivity that is inherent in evaluation and in every human endeavor. She argues that standards and guidelines for ethical conduct cannot anticipate the wide range of particularities that are present in any evaluation. Thus, evaluators' personal standards and judgment inevitably play a role in how they apply these codes of conduct to the evaluations they carry out.

Perhaps Sieber still states it best:

> A code of ethics specifically for program evaluators . . . would be a minimum standard; it would only state what the profession expects of every evaluator in the way of honesty, competence, and decency in relation to those ethical problems that are clearly defined at present.

> In contrast, being ethical is a broad, evolving personal process. . . . Ethical problems in program evaluation are problems having to do with unanticipated conflicts of obligation and interest and with unintended harmful side effects of evaluation. To be ethical is to evolve an ability to anticipate and circumvent such problems. It is an acquired ability. . . . As one undertakes new and different kinds of evaluation and as society changes, one's ability to be ethical must grow to meet new challenges. Thus, being ethical in program evaluation is a process of growth in understanding, perception, and creative problem-solving ability that respects the interests of individuals and of society. (1980, p. 53)

Major Concepts and Theories

1. Good evaluation practice involves much more than methodological skills. Evaluators must have the skills to work well in a sometimes highly political environment, must be able to communicate well with clients and other stakeholder groups, and must know the ethical problems evaluators can encounter and the ethical expectations for good evaluation practice.

2. Evaluations are an inherently political activity because they are concerned with guiding or influencing public policies, because their results can have powerful implications for individuals and stakeholder groups competing for power, and because they concern human beings, organizations, and judgments about programs.

3. Evaluators need to have skills to work in a political environment both to increase the likelihood that the evaluation will be used and to prevent political actions that may bias the results. These skills include learning about the political environment and the positions of the stakeholders in it, considering and including other stakeholders and the public in the evaluation as appropriate, and working to maintain the credibility of the evaluation.

4. Evaluators should foster good communication with stakeholders by listening to their concerns and learning about their experiences with evaluation, educating stakeholders on the different purposes evaluation can serve, meeting frequently with the client and other appropriate stakeholders, and involving them in the decisions made concerning the evaluation.

5. The *Program Evaluation Standards* and the American Evaluation Association's Guiding Principles provide guidance for the conduct of good and ethical evaluations. Evaluators should be knowledgeable about the standards and principles for their country and use them to inform clients and other stakeholders of the expectations for them as a professional.

6. Protecting the rights of those who provide data for the study is essential to good, ethical evaluation. Such rights include having a free choice to participate without threat of losing services, understanding the nature of the evaluation and the data collection and its potential risks and benefits, being informed about confidentiality and its limitations, and being treated with respect and dignity. Evaluators should seek the input or approval of Institutional Review Boards or others informed on the ethics of data collection to ensure that appropriate precautions are taken.

7. Evaluations can be biased by the personal views and experiences of the evaluator; by his or her views and relationships with program staff, administrators, and clients; or by financial or organizational pressures. The evaluator should be conscious of these sources of bias and seek to avoid relationships that would unduly threaten the perceived neutrality of the evaluation findings. The evaluator should work to gain cultural competence in the setting of the evaluation and consider the cultural views of others.

8. Ethical practice requires evaluators not only to become familiar with the *Standards* and the Guiding Principles and to acquaint clients with these professional expectations, but also to carefully consider decisions throughout the evaluation in terms of potential ethical concerns. Professional codes can be one source for resolution of ethical problems, but continued personal growth, reading, reflection, and discussion with others are essential.

Discussion Questions

1. What are the good elements of evaluation studies taking place in a political environment? The bad elements? How does politics enter into evaluations that you know about?

2. Which of the three positions described by Vestman and Conner for evaluators to take in a political environment do you feel are most appropriate? Why?

3. Why is there a need for explicit ethical standards in evaluation? What benefits accrue to the evaluator and client by adhering to these standards?

4. What types of ethical violations do you think would occur most commonly in organizations with which you are familiar? How might these violations be prevented?

Application Exercises

For exercises 1 to 3, consider an evaluation in which you were a participant or the evaluator.

1. How did politics enter into this evaluation? Did the politics introduce bias or problems? How did the evaluator attend to the political context?

2. How did the evaluator or the evaluation team communicate with you and other key stakeholders? On what issues did they seek your input? Do you think the relationships the evaluator established with you and other stakeholders in the evaluation led to bias, or did it improve the evaluation?

3. Consider this evaluation in reference to the *Program Evaluation Standards* and the AEA Guiding Principles. What were the ethical strengths and weaknesses of the evaluation?

4. Now consider a program that you are familiar with—perhaps one in your organization. If you had to evaluate that program, what biases would you bring? Do you think you would be an appropriate person to evaluate it? Who (person or organization) might be the best alternative? Why?

Case Studies

To close this chapter on politics, interpersonal relationships, and ethics, we resume the practice begun in Chapter 1 of recommending interviews that describe an evaluation that illustrates the issues discussed in the chapter. The interviews we recommend for this chapter are in *Evaluation in Action,* Chapters 4 (Len Bickman) and 12 (Katrina Bledsoe).

In Chapter 4, Len Bickman, a past president of the American Evaluation Association, describes some of the difficult political circumstances he encountered in a nationally recognized evaluation of mental health systems of care. The journal source is Fitzpatrick, J. L., & Bickman, L. (2002). Evaluation of the Ft. Bragg and Stark County systems of care for children and adolescents: A dialogue with Len Bickman. *American Journal of Evaluation, 23,* 67–80.

In Chapter 12, Katrina Bledsoe describes her evaluation of a program for parents and preschool children to encourage reading and preliteracy skills. She demonstrates her skills at developing strong interpersonal relationships with people in the program, working at achieving cultural competence and understanding different cultural views, and facing ethical challenges from the client on the final report. The journal source is Fitzpatrick, J. L., & Bledsoe, K. (2007). Evaluation of the Fun with Books Program: A dialogue with Katrina Bledsoe. *American Journal of Evaluation, 28,* 522–535.

Suggested Readings

Chelimsky, E. (2009). A clash of cultures: Improving the "fit" between evaluative independence and the political requirements of a democratic society. *American Journal of Evaluation, 29,* 400–415.

Joint Committee on Standards for Educational Evaluation. (2010). *The program evaluation standards.* Thousand Oaks, CA: Sage.

Morris, M. (Ed.). (2008). *Evaluation ethics for best practice: Cases and commentary.* New York: Guilford Press.

Shadish, W. R., Newman, D. L., Scheirer, M. A., & Wye, C. (Eds.). (1995). *Guiding principles for evaluators.* New Directions for Program Evaluation, No. 66. San Francisco: Jossey-Bass.

Vestman, O. K., & Conner, R. F. (2006). The relationship between evaluation and politics. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation.* Thousand Oaks, CA: Sage.

This page intentionally left blank

Part II

Alternative Approaches to Program Evaluation

In Part One, we referred to the varying roles that evaluation studies can play in education, government, business, nonprofit agencies, and many related areas, and readers were introduced to some of the different purposes of evaluation. We hinted at some of the different approaches to evaluation, but we have not yet exposed the reader to these approaches. We will do so in Part Two.

In Chapter 4, we examine the factors that have contributed to such differing views. Prior efforts to classify the many evaluation approaches into fewer categories are discussed, and the categories that we will use in the remainder of this book are presented.

In Chapters 5 through 8, we describe four categories of approaches that have influenced evaluation practice. These general approaches include those we see as most prevalent in the literature and most popular in use. Within each chapter, we discuss how this category of approaches emerged in evaluation, its primary characteristics, and how it is used today. Within some categories, there are several major approaches. For example, participatory evaluation has many models or approaches. We describe each approach, including its distinguishing characteristics and contributions, the ways in which the approach has been used, and its strengths and weaknesses. Then, in Chapter 9, we discuss other themes or movements in evaluation that transcend individual models or approaches, but that are important influences on evaluation practice today.

Many evaluation books, often authored by the developer of one of the approaches we discuss, present what Alkin (2004) has called "prescriptive theories" or approaches to evaluation. These books are intended to describe that approach in depth and, in fact, to suggest that the approach presented is the one that evaluators should follow. This book does not advocate a particular approach. Instead, we think it is important for evaluators and students studying evaluation to be familiar with the different approaches so they can make informed choices concerning

which approach or which parts of various approaches to use in a particular evaluation. Each approach we describe tells us something about evaluation, perspectives we might take, and how we might carry out the evaluation. During this time of increased demands for evaluation in the United States and the world—what Donaldson and Scriven (2003) have called the "second boom in evaluation"—it is important for evaluators to be aware of the entire array of evaluation approaches and to select the elements that are most appropriate for the program they are evaluating, the needs of clients and other stakeholders, and the context of the evaluation.

4

Alternative Views of Evaluation

Orienting Questions

1. Why are there so many different approaches to evaluation?
2. Why is evaluation theory, as reflected in different approaches to evaluation, important to learn?
3. What philosophical and methodological differences influenced the development of different approaches?
4. How have evaluation approaches been categorized by others? How does this book categorize evaluation approaches? What is the rationale for each?
5. What practical issues contribute to the diversity of evaluation approaches?

In the early days, when evaluation was emerging as a field, it was troubled by definitional and ideological disputes. Those who wrote about evaluation differed widely in their views of what evaluation was, and those who conducted evaluation studies brought to the task diverse conceptions of how one should go about doing it. From 1960 to 1990, nearly 60 different proposals for how evaluations should be conducted were developed and circulated. These proposals have been chronicled from the early days of thinking about evaluation approaches (Gephart, 1978) to more recent reviews of the development of evaluation models (Stufflebeam, 2001b). These different prescriptions have been implemented with varying degrees of fidelity. To complicate the picture further, some evaluations were designed without conscious reference to any existing conceptual framework, occasionally resulting, if successful, in yet another evaluation approach.

The various approaches, or theories, proposed by evaluators make up the content of the field of evaluation. William Shadish titled his presidential address to the American Evaluation Association in 1997 "Evaluation Theory Is Who We Are" and argued that "[a]ll evaluators should know evaluation theory because it is central to our professional identity" (1998, p. 1). As he pointed out, evaluation theory "provides the language we use to talk to each other about evaluation" and "is the knowledge base that defines the profession" (Shadish, 1998, pp. 3, 5). Stufflebeam, too, emphasizes the importance of studying evaluation theory and its approaches. He writes, "The study of alternative evaluation approaches is important for professionalizing program evaluation and for its scientific advancement and operation" (2001b, p. 9). As illustrated in Shadish's and Stufflebeam's remarks, some evaluators use the term evaluation "theories"; others use the terms evaluation "models" or "approaches." We prefer to use the word approaches, because few are as broad as a true theory and their intent is to guide how evaluation is practiced.[1]

Today, although there is no dominant evaluation theory or approach, there is much more agreement than in the past. Nevertheless, it is important for readers to become familiar with the different approaches, not only to learn the knowledge base of the field and the issues that professional evaluators discuss, but also to help them make conscious choices about the approach or elements of different approaches that they intend to use in each evaluation. Many evaluators today use a mix of approaches, selecting elements that are most appropriate for the program they are evaluating, its context, and stakeholders. Sometimes a funder will select the approach to be used, although evaluators may choose to negotiate changes to that if the funder is not familiar with other approaches and the one chosen is inappropriate for the program or its context. But, without knowledge of these different approaches, evaluators tend to make uninformed choices of the questions their evaluation should address, the ways in which stakeholders might be involved, the appropriate methods to use for collecting data, and the means for maximizing the use of the results. (See, for example, Christie's 2003 study of practicing evaluators.)

Approaches to evaluation that have emerged as the most common or well-known are described in the chapters following Chapter 4. These approaches provide the conceptual tools for an evaluator to use in designing an evaluation that fits particular circumstances. In this chapter, we will discuss the factors that have influenced the differences in approaches, some of the ways in which approaches have been categorized, and how we have conceptualized the common approaches used today.

[1]Shadish (1998) defines "theory" in his address as "a whole host of more or less theoretical writings with evaluation as their primary focus" (p. 1). Like "approaches," these writings discuss how evaluation should be conducted and the factors that influence its practice.

Diverse Conceptions of Program Evaluation

The many evaluation approaches that have emerged since 1960 range from comprehensive models to checklists of actions to be taken. Some authors opt for a comprehensive approach to judging a program, while others view evaluation as a process of identifying and collecting information to assist decision makers. Still others see evaluation as synonymous with professional judgment, where judgments about a program's quality are based on opinions of experts. In one school of thought, evaluation is viewed as the process of comparing performance data with clearly specified goals or objectives, while in another, it is seen as synonymous with carefully controlled experimental research on programs to establish causal links between programs and outcomes. Some focus on the importance of naturalistic inquiry or urge that value pluralism be recognized, accommodated, and preserved. Others focus on social equity and argue that those involved with the entity being evaluated should play an important, or even the primary, role in determining what direction the evaluation study takes and how it is conducted.

The various models are built on differing—often conflicting—conceptions and definitions of evaluation. Let us consider an example from education.

- If one viewed evaluation as essentially synonymous with professional judgment, the worth of an educational program would be assessed by experts (often in the subject matter to be studied) who observed the program in action, examined the curriculum materials, or in some other way gleaned sufficient information to record their considered judgments.
- If evaluation is viewed as a comparison between student performance indicators and objectives, standards would be established for the curriculum and relevant student knowledge or skills would be measured against this yardstick, using either standardized or evaluator-constructed instruments.
- If an evaluation is viewed as providing useful information for decision making, the evaluator, working closely with the decision maker(s), would identify the decisions to be made and collect sufficient information about the relative advantages and disadvantages of each decision alternative to judge which was best. Or, if the decision alternatives were more ambiguous, the evaluator might collect information to help define or analyze the decisions to be made.
- If the evaluator emphasized a participative approach, he or she would identify the relevant stakeholder groups and seek information on their views of the program and, possibly, their information needs. The data collection would focus on qualitative measures, such as interviews, observations, and content analysis of documents, designed to provide multiple perspectives on the program. Stakeholders might be involved at each stage of the evaluation to help build evaluation capacity and to ensure that the methods used, the interpretation of the results, and the final conclusions reflected the multiple perspectives of the stakeholders.

- If the evaluator saw evaluation as critical for establishing the causal links between the program activities and outcomes, he or she might use random assignment of students, teachers, or schools to the program and its alternatives; collect quantitative data on the intended outcomes; and draw conclusions about the program's success in achieving those outcomes.

As these examples illustrate, the way in which one views evaluation has a direct impact on the manner in which the evaluation is planned and the types of evaluation methods that are used. Each of the previous examples, when reviewed in detail, might be considered an excellent evaluation. But, evaluations must consider the context in which they are to be conducted and used. Each context—the nature and stage of the program, the primary audiences for the study and the needs and expectations of other stakeholders, and the political environment in which the program operates—holds clues to the approach that will be most appropriate for conducting an evaluation study that makes a difference in that context. Therefore, without a description of the context, we cannot even begin to consider which of the examples would lead to the best evaluation study. Nor can we judge, based on our own values, which example is most appropriate. Instead, we must learn about the characteristics and critical factors of each approach so that we can make appropriate choices when conducting an evaluation in a specific context.

Origins of Alternative Views of Evaluation

The diversity of evaluation approaches has arisen from the varied backgrounds, experiences, and worldviews of their authors, which have resulted in diverse philosophical orientations, and methodological and practical preferences. These different predispositions have led the authors—and their adherents—to propose sometimes widely different methods for conducting evaluations and for collecting and interpreting information or data. The differences in evaluation approaches can be traced directly to their proponents' rather different views not only of the meaning and nature of evaluation but also of the nature of reality (ontology) and knowledge (epistemology).

To understand the origins of alternative conceptualizations of evaluation, the reader will first need an introduction to different philosophical views of ontology and epistemology.

Philosophical and Ideological Differences

Logical Positivism. Early evaluations emerged from the social sciences, in particular education and psychology, at a time when the dominant paradigm was positivism. Logical positivists, a more extreme branch of positivism, argued that knowledge was obtained entirely through experience, specifically through observation, and held rigid views concerning the world and data collection (Godfrey-Smith, 2003). They argued that (a) there is one reality of the objects we are studying and the aim

of researchers and evaluators is to use social science research methods and theories of statistical probability to discover that one reality and to establish laws and theories about how things work, and (b) to effectively gain knowledge of that reality, researchers need to be "scientifically objective." A key component of that approach is that researchers should maintain some distance from the program to be studied so as not to influence the program itself, the participants, or the results of the study. The methods used to achieve this objectivity, or distance, were typically quantitative in nature. Objectivity or objectivism, meaning that the researcher's views and values do not influence the results obtained, was a key principle of positivism.

Postpositivism. Reichardt and Rallis (1994) note that logical positivism began to decline around the time of World War II, though elements of positivism continued to influence research and evaluation for some time. By 1984, however, Donald Campbell, a prominent research methodologist and evaluator with a quantitative orientation, noted that "twenty years ago logical positivism dominated the philosophy of science. . . . Today the tide has completely turned among the theorists of science in philosophy, sociology, and elsewhere. Logical positivism is almost universally rejected" (p. 27). Postpositivism emerged in reaction to logical positivism and many, unfortunately, confuse the two. Guba and Lincoln (1989) argued that the views of postpositivists were not compatible with other approaches to evaluation. However, Reichardt and Rallis (1994), quantitative and qualitative evaluators respectively, effectively refuted their arguments, demonstrating that postpositivists, such as Campbell and Stanley (1966) and Cook and Campbell (1979), did not hold the views of logical positivists. Instead, they showed through quotations from their work that these postpositivists, and others, recognized that facts and methods or inquiry choices in research are influenced by the values of the researcher, that knowledge is fallible and changing, that data can be explained by many different theories, and that reality is constructed by people and their experiences.

The focus of postpositivists, however, was on examining causal relationships to develop laws and theories to describe the external world, albeit temporary ones given the fallibility of knowledge. Replication and intersubjectivity, not objectivity, were the keys to ensuring good research (Frankfort-Nachmias & Nachmias, 2008). Intersubjectivity involves the ability to communicate what one does in research in such a way that others can judge its findings and replicate it to see if they obtain the same results. For evaluation, House and Howe (1999) note that one of the key characteristics of this philosophical approach, which they call the received view, is viewing facts as quite distinct from values and believing that evaluators should be focusing on the facts.

A Constructivist Paradigm. As evaluation continued, evaluators saw that context and values played very important roles in evaluation. Unlike many laws of science which are readily generalizable from one setting to the next, the factors that influence the success of education, social, and economic programs can differ dramatically from one setting to another. Also, clients and stakeholders for the

evaluation often had information needs that were not so concerned with establishing causality as with gaining a better understanding of the program and those they served. Program developers recognized the many differing "realities" or conditions or life experiences of those that the programs were intended to serve and saw that programs had different effects on different kinds of clients. They wanted to know more about these issues to help them improve their programs. And values were an integral part of what programs, policies, and evaluations confronted. To exempt evaluation from such values was to make it incomplete.

The constructivist paradigm that was emerging then corresponded more closely to the views and experiences of these evaluators and program developers. Constructivists took a different view of ontology and epistemology (Guba & Lincoln, 1994). Although we now realize that the differences were not as extreme as they were sometimes portrayed, Guba and Lincoln focused on *understanding* our constructed world and, in particular, the multiple realities seen or experienced by different stakeholders. They argued that objectivity was not possible; we each see the world through our own lens, influenced by our own experiences. Later, House and Howe (1999) emphasized that the fact-value dichotomy, or the rigid distinction between "facts" which are objective and "values" which are subjective, is in fact (pun intended) a continuum. Our values influence what we perceive to be facts. Thus, evaluators should become involved with values—helping stakeholders articulate their values, considering the values inherent in the evaluation, and working to portray the program through different stakeholders' perspectives of reality. Constructivism also continued its focus on what Schwandt (1997) calls the "localness" of knowledge. Evaluation is intended to provide understanding of a particular program and its context and is less concerned with generalizability and developing laws and theories for other settings.

A Transformative Paradigm. More recently, a new paradigm for evaluation has emerged—the transformative paradigm. It emerged initially, and is still most powerful, in international development work and in the developing world, though the paradigm is gaining proponents in the United States and Western countries. Like constructivism and postpositivism, this paradigm emerged in response to the strictures of positivism, but also developed in response to concerns in developing countries that research and evaluation often failed to address critical political and social problems. Like the constructivist paradigm, the transformative paradigm acknowledges multiple realities and the need for evaluation to capture those realities. However, the emphasis of the transformative paradigm is on the political, social, and economic factors that form those realities. The transformative paradigm is less concerned with methodological choices and more concerned with the nature of the problems that evaluation addresses and how stakeholders are involved in the evaluation. Transformative evaluations are concerned with empowering groups that have less power in society. These can include poor people, ethnic or racial minorities, women, and people with disabilities (Mertens, 1999). The focus of the evaluation is on helping these groups construct their own knowledge and

empowering them by having them play a central role in the evaluation (Hall, 1992; Freire, 1970, 1982). The evaluator serves as a facilitator to the decisions made by the stakeholders about the evaluation in order to change power structures and knowledge. Some view transformative evaluation as a new paradigm. Others view it as an approach. We will cover this type of evaluation as an approach more extensively in Chapter 8.

The Influence of Paradigms on Evaluation Practice. These philosophical paradigms, and their implications for methodological choices, have influenced the development of different evaluation approaches. Some have argued that paradigms and qualitative and quantitative methods should not be mixed because the core beliefs of postpositivists and constructivists are incompatible (Denzin & Lincoln, 1994). As noted, Reichardt and Rallis (1994) argued and demonstrated that the paradigms were compatible. These and other pragmatists, representing different methodological stances—quantitative and qualitative—disputed the incompatibility argument and urged evaluators and researchers to look beyond ontological and epistemological arguments to consider what they are studying and the appropriate methods for studying the issues of concern. In other words, evaluative and methodological choices should not be based on paradigms or philosophical views, but on the practical characteristics of each specific evaluation and the concepts to be measured in that particular study. Today, there are many evaluators, some of whose approaches will be discussed in subsequent chapters, who skip the arguments over paradigms and prefer a pragmatic approach (Patton, 1990; 2001; Tashakkori and Teddlie, 2003). Howe (1988) and, more recently, Tashakkori and Teddlie (1998) have proposed the pragmatic approach as a paradigm in itself. They see discussions of ontology and epistemology as fruitless and unnecessary and argue that researchers' and evaluators' choice of methods should be based on the questions the evaluator or researcher is trying to answer. They write, "Pragmatist researchers consider the research question to be more important than either the methods they use or the paradigm that underlies the method" (Tashakkori & Teddlie, p. 21, 2003).

It is useful, however, for readers to be familiar with these paradigms because their philosophical assumptions were key influences on the development of different evaluation approaches and continue to play a role in many evaluations and approaches.

Methodological Backgrounds and Preferences

For many years, evaluators differed, and argued, about the use and value of qualitative or quantitative methods, as suggested previously. These methodological preferences were derived from the older paradigms described earlier. That is, the postpositivist paradigm focused on quantitative methods as a better way to obtain objective information about causal relationships among the phenomena that evaluators and researchers studied. To be clear, quantitative methods are ones that yield numerical data. These may include tests, surveys, and direct measures of certain quantifiable constructs such as the percentage of entering students who

graduate from a high school to examine a school's success, blood alcohol content for the evaluation of a drunk-drivers treatment program, or the numbers of people who are unemployed to evaluate economic development programs. Quantitative methods also rely on experimental and quasi-experimental designs, or multivariate statistical methods, to establish causality.

Constructivists were more concerned with describing different perspectives and with exploring and discovering new theories. Guba and Lincoln discussed developing "thick descriptions" of the phenomenon being studied. Such in-depth descriptions were more likely to be made using qualitative observations, interviews, and analyses of existing documents. Constructivists also see the benefit of studying causal relationships, but their emphasis is more on understanding those causal relationships than on establishing a definitive causal link between a program and an outcome. Given these emphases, constructivists favored qualitative measures. Qualitative measures are not readily reducible to numbers and include data collection methods such as interviews, focus groups, observations, and content analysis of existing documents.

Some evaluators have noted that the quantitative approach is often used for theory testing or confirmation while qualitative approaches are often used for exploration and theory development (Sechrest & Figueredo, 1993; Tashakkori & Teddlie, 1998). If the program to be evaluated is based on an established theory and the interest of the evaluation is in determining whether that theory applies in this new setting, a quantitative approach might be used to determine if, in fact, the causal mechanisms or effects hypothesized by the theory actually did occur. For example, a reading program based upon an established theory is being tried with a younger age group or in a new school setting. The focus is on determining whether the theory works in this new setting to increase reading comprehension as it has in other settings. Students might be randomly assigned to either the new method or the old one for a period of a few months, and then data would be collected through tests of reading comprehension. While qualitative methods could also be used to examine the causal connections, if the focus were on firmly establishing causality, quantitative approaches might be preferred. In contrast, if the evaluator is evaluating an experimental program or policy for which the theory is only loosely developed—for example, a new merit pay program for teachers in a particular school district—a qualitative approach would generally be more appropriate to better describe and understand what is going on in the program. Although a few districts are experimenting today with merit pay, little is known about how merit pay might work in educational settings, and results from other sectors are mixed (Perry, Engbers, & Jun, 2009; Springer & Winters, 2009). In this case, it would be important to collect much qualitative data through interviews with teachers, principals, and other staff; observations at staff meetings; content analysis of policy documents; and other methods to learn more about the impact of merit pay on the school environment; teacher retention, satisfaction, and performance; teamwork; teacher-principal relations; and many other issues.

In the beginning years of evaluation, most evaluators' training was in quantitative methods. This was particularly true for evaluators coming from the disciplines

of psychology, education, and sociology. The emergence of qualitative methods in evaluation provided new methodologies that were initially resisted by those more accustomed to quantitative measures. Today, however, most evaluators (and researchers) acknowledge the value of mixed methods and most graduate programs recognize the need to train their students in each, though some may focus more on one method than another. For researchers, who tend to study the same or a similar subject most of their career, intensive training in a few methodologies appropriate for the types of constructs and settings they are studying is appropriate. But evaluators study many different programs and policies containing many different important constructs over the course of their careers. Therefore, evaluators now recognize the need to have skills in both qualitative and quantitative methods in order to select the most appropriate method for the program and context they are evaluating.

One useful framework for explaining the differences among evaluators and approaches over the years comes from Stevenson and Thomas (2006), who analyzed what they called the intellectual contexts for evaluation. They identified three traditions in evaluation that are closely tied to one's original training and discipline:

(a) The *experimental tradition* is composed primarily of people trained in psychology and sociology, and in quantitative methods with a focus on establishing causality. Donald Campbell was an early leader in this tradition, moving social psychologists to think more practically about conducting useful research beyond the laboratory.
(b) The *case/context tradition,* led by Ralph Tyler and his student Lee Cronbach, is primarily grounded in education. This movement was rooted in testing and student assessment, but moved on to describe programs and work with teachers to gain an understanding of what was happening.
(c) The *policy influence tradition* is composed of people trained in political science and often working in the federal government. These leaders included Carol Weiss and Joseph Wholey. Their work on policy, which was somewhat removed from individual programs but tried to help elected and appointed government officials make decisions about what to fund and the directions government should take, led to a different kind of focus on use and designs.

Although evaluators come together today at large meetings of professional associations, such as the American Evaluation Association attended by more than 2,000 evaluators, these traditions can still be seen. They learn a little from each other, but continue, often, to focus on the issues familiar to the environments in which they work and their original training. By presenting different approaches in this textbook, we hope to help readers bridge these disciplines and traditions and learn what might be valuable from each for the context in which they work.

Disciplinary Boundaries and Evaluation Methodology. It is ironic that in a field with such a rich array of alternative evaluation approaches, there still exists, among some evaluators, a tendency to fall prey to the law of the instrument

fallacy[2] rather than to adapt or develop evaluation methods to meet the needs of the program, the stakeholders, and the identified evaluation questions. In many cases, the law of the instrument fallacy in evaluation is grounded in the methods of the discipline of the evaluator's original training. However, Scriven (1991c) has effectively argued that evaluation is not a single discipline but a transdiscipline that, like logic, design, and statistics, is applied to a wide range of disciplines.

Thus, our presentation of approaches is not meant to encourage a single approach, but to encourage the reader to adopt the approach or elements of different approaches that are appropriate for the particular evaluation he or she is planning.

Classifications of Evaluation Theories or Approaches

Existing Categories and Critiques

In recent years, several evaluators have attempted to categorize evaluation theories for different purposes. Shadish, Cook, and Leviton's (1995) book was influential in reviewing important evaluation theorists, at least partly to illustrate historical trends and changes in the field, but primarily to identify and describe important evaluation theories. Shadish et al. identified three stages of evaluation theory as it emerged in the United States. The first stage, in the 1960s, was characterized by a focus on using scientifically rigorous evaluation methods for social problem solving or studying the effectiveness of government programs at achieving outcomes. The emphasis at this stage of evaluation was on examining causal effects of programs and, with this information, judging the value of each program. Shadish et al. focus on individual evaluators to illustrate the dominant theories at each stage. For the first stage, they profile Michael Scriven, who developed his theory of valuing—the process of reaching a judgment on the value of programs or policies—and Donald Campbell, who developed quasi-experimental methods to establish the causal effects of programs outside of the laboratory and discussed how these methods should be used by managers and evaluators. Stage two reflected evaluators' growing concern with having evaluation results used.[3] Evaluators' focus on use in stage two prompted evaluation to grow and change in many ways, such as encouraging evaluators to establish relationships with specific stakeholders to facilitate use, and broadening the methods used to accommodate

[2]Kaplan (1964) described this fallacy by noting that, if you give a small boy a hammer, suddenly everything he encounters needs hammering. The same tendency is true, he asserts, for scientists who gain familiarity and comfort in using a particular method or technique: suddenly all problems will be wrested into a form so that they can be addressed in that fashion, whether or not it is appropriate.

[3]Stage one theorists had not written extensively about use, assuming results would naturally be used by consumers, managers, or policymakers.

the potential information needs and values of different stakeholders. The theorists they profile in the second stage, Carol Weiss, Joseph Wholey, and Robert Stake, were concerned, in quite different ways, with increasing the responsiveness and utility of evaluations. In stage three, Shadish et al. view evaluators such as Lee Cronbach and Peter Rossi as integrating the first stage's emphasis on truth or scientific validity with the second stage's emphasis on evaluation's utility to stakeholders. In efforts to have evaluation be both valid and useful, stage three evaluators introduce new concepts such as developing the theory of a social program to aid in its evaluation and adapt others.

Stufflebeam (2001b), too, analyzed evaluation theories or, what he, like us, calls "approaches." His work was designed to reduce the burgeoning numbers of evaluation theories and to identify those with the greatest potential. He attempted to reduce the numbers of theories to those that are most useful by conducting an intensive study of 20 different evaluation approaches using some key descriptors to summarize each approach. He then used the *Standards* developed by the Joint Committee to judge nine approaches in more detail. His assessments of the various methods were also influenced by the extent to which each approach addresses what he sees as "evaluation's fundamental requirement to assess a program's merit or worth" (Stufflebeam, 2001b, p. 42). Of interest to us here is his categorization of the 20 approaches into three groups: (a) Question and/or Methods-Oriented approaches, (b) Improvement/Accountability approaches, and (c) Social Agenda/Advocacy approaches. His first category, Question and/or Methods-Oriented approaches, is the largest of the three groups, containing 13 of the 20 approaches. These approaches, Stufflebeam notes, are alike in that they "tend to narrow an evaluation's scope" by focusing on either particular questions or methods (2001b, p. 16). Approaches in this category include ones that focus on particular strategies to determine what should be evaluated (objectives-oriented and theory-based approaches), on particular methods to collect data (objective testing, performance testing, experimental studies, case studies, cost-benefit analysis) or to organize data (management information systems), or on a particular method for presenting and judging results (clarification hearing).[4] Stufflebeam's second category, Improvement/Accountability approaches, contains approaches that"stress the need to fully assess a program's merit or worth" (2001b, p. 42). Stufflebeam sees these approaches as more comprehensive in their evaluation of programs in order to serve their purpose of judging merit or worth. Typical examples include the accreditation/certification approach and Scriven's consumer-oriented approach to judging the quality of products for potential consumers. The Social Agenda/Advocacy approaches, rather than having a primary emphasis on judging the overall quality of a product or relying upon a particular method, "are directed to making a difference in society through program evaluation" (Stufflebeam, 2001b, p. 62). In conducting evaluations, these approaches are concerned with involving or empowering groups who have less power in society. These

[4]These sub-categories are our own interpretation of the 13 approaches, not Stufflebeam's.

approaches include Stake's client-centered or responsive evaluation and House's deliberative democratic evaluation.

In 1985, Alkin and Ellett argued that to be considered a comprehensive theory, evaluation theories must address three issues: methodologies, how the data are valued or judged, and use of the evaluation. Later Alkin and House (1992) developed these issues into three continua: (a) methods could be characterized along a continuum from qualitative to quantitative; (b) values could be characterized from unitary (one value or way of judging the data and program) to plural (many values); and (c) use could range from aspirations for instrumental, or direct use, to enlightenment or indirect use. In 2004, Alkin and Christie used these dimensions to categorize evaluation theorists and their approaches through the visual model of a tree. The roots of the tree reflect what they see as the dual foundations of evaluation: social inquiry (using a "systematic and justifiable set of methods") and accountability and control (reflecting the purposes and intended use of evaluation). The branches of the tree then reflect the three dimensions of methods, values, and use identified earlier by Alkin and House (1992). Individual theorists are placed on one of the three branches to reflect the key dimension of their approaches. Like Shadish et al. (1995), Alkin and Christie use individual evaluation theorists to illustrate different approaches to evaluation.

Each of these categorizations of evaluation approaches or theories provides useful insights into evaluation and its history and practice. Thus, Shadish, Cook, and Leviton illustrate the early focus on the truth that evaluation would bring to the judgments made about social programs, the later recognition that use needed to be consciously considered, and the integration and adaptation of the two issues in even later stages. Alkin and Christie's model builds on these foundations identified by Shadish et al. in slightly different ways. Its roots are in social inquiry, accountability, and control, but it considers evaluation's emphases in three areas: methods, values, and use. Stufflebeam's categories are different from the first two in that he focuses not on individual evaluators and their writings to identify categories, but on the content of evaluation theories or models.[5] He developed his categories by considering the orienting devices or principles used for focusing the evaluation. The priorities used to focus the evaluation are reflected in his three categories: using particular evaluation questions or methods, taking a comprehensive approach to making a judgment regarding quality of the program, or improving society and its programs by considering social equity and the needs of those with less power. Like Stufflebeam, our purpose is to reduce the current number of evaluation approaches. Although Stufflebeam's method of reducing the approaches was to judge the quality of each, our synthesis of the approaches is intended to *describe* each approach to help you, the reader, to consider different

[5]Of course, examining the writings of proponents leads one to consider the theories as well because the individual is writing about his or her evaluation approach or theory. The difference is that Alkin and Christie (2004) and Shadish et al. (1995) were focusing on individuals to illustrate theories, and Stufflebeam's writing is less concerned with individuals. Although some of the theories Stufflebeam reviews are identified with one individual, others are not.

approaches and their potential use in your work.[6] Although the many different approaches to evaluation can seem confusing, their diversity allows evaluators to pick and choose either the approach or the elements of an approach that work best for each program they are evaluating. Our task is to categorize the approaches in a way that helps you consider them and to expand your views of possible ways in which evaluations may be conducted.

A Classification Schema for Evaluation Approaches

We have chosen to classify the many different evaluation approaches into the four categories that we have developed based on our identification of the primary factor that guides or directs the evaluation:

1. *Approaches oriented to comprehensive judgments of the quality of the program or product:* These approaches include expertise-oriented and consumer-oriented evaluations. They are the oldest approaches in evaluation, having been used by many before formal evaluation approaches were developed. We will discuss Elliot Eisner's writings on connoisseurship and criticism, accreditation, and other types of expertise-oriented evaluations and Michael Scriven's consumer-oriented approach. The expertise and consumer-oriented approaches differ rather dramatically in who conducts the evaluation and the methodology, but their commonality is that they direct evaluators to focus on valuing or judging the quality of the thing they are evaluating.
2. *Approaches oriented to characteristics of the program:* These approaches include objectives-based, standards-based, and theory-based evaluations. In each of these approaches, the evaluator uses characteristics of the program, its objectives, the standards it is designed to achieve, or the theory on which the program is based to identify which evaluation questions will be the focus of the evaluation.
3. *Approaches oriented to decisions to be made about the program:* These approaches include Daniel Stufflebeam's Context-Input-Process-Product (CIPP) approach and Michael Patton's Utilization-Focused Evaluation, as well as Joseph Wholey's evaluability assessment and performance monitoring. These approaches focus on evaluation's role in providing information to improve the quality of decisions made by stakeholders or organizations.
4. *Approaches oriented to participation of stakeholders:* These approaches include Robert Stake's Responsive Evaluation, Practical Participatory Evaluation, Developmental Evaluation, Empowerment Evaluation, and democratically oriented approaches.

Placement of individual evaluation approaches within these categories is to some degree arbitrary. Several approaches are multifaceted and include characteristics that would allow them to be placed in more than one category. For clarity we have

[6]Although our purpose is not to judge the quality of each approach, but to introduce you to them, we do not include approaches that could not serve a valid purpose in an evaluation.

decided to place such approaches in one category and only reference their other features in chapters where it is appropriate. Our classification is based on what we see as the driving force behind doing the evaluation: the factors that influence the choice of what to study and the ways in which the study is conducted. Within each category, the approaches vary by level of formality and structure, some being relatively well developed philosophically and procedurally, others less developed. Some are used frequently; others are used less, but have had a major influence on evaluators' thinking. The following chapters will expand on these approaches.

Major Concepts and Theories

1. During evaluation's relatively short history, many different approaches or theories concerning how to practice evaluation have emerged.

2. Evaluators should be familiar with the various approaches in order to choose the approach or elements of approaches most appropriate for the specific program they are evaluating and its context.

3. The different evaluation approaches were influenced by differing views of ontology (the world and reality) and epistemology (knowledge), and the methods for obtaining valid knowledge. These views often are associated with the evaluator's graduate training and life experiences.

4. Today, prominent paradigms in evaluation and the social sciences include postpositivist, constructivist, transformative, and pragmatist paradigms.

5. Others have categorized evaluation theories or approaches according to a focus on truth and use and an integration of the two; by the categories of questions or methods, improvement/accountability, and social agenda/advocacy; and by their focus on methods, values, or use.

6. We categorize theories based on the primary factor that guides the actions taken in the evaluation. Our categories include approaches that focus on making an overall judgment regarding the quality of the program, on program characteristics, on decisions to be made, and on stakeholder participation.

Discussion Questions

1. What are the key differences between the paradigms that have influenced evaluation? Which paradigm seems most appropriate to you? Why?

2. How can the ways in which one defines program evaluation impact an evaluation study?

3. What implications does the statement "evaluation is not a traditional discipline but a transdiscipline" have for the methodologies or approaches an evaluator may decide to use in an evaluation?

Application Exercises

1. Think about how you could approach evaluation. Describe the steps you think you would follow. Then, analyze your approach according to your philosophical and methodological preferences. Explain how your background and what you would be evaluating could affect your approach. Describe other things that might affect your approach to evaluation.
2. Identify a program in your area that you would like to see evaluated. List some qualitative evaluation methods that could be used. Now list some quantitative methods that you see as appropriate. What might the different methods tell you?
3. The Anderson Public School District has recently begun a new training program for principals. What questions would you ask if you were to conduct an evaluation of this training program from the postpositivist paradigm? What types of data would you collect? How might this evaluation be conducted differently if you took a constructivist perspective? A transformative perspective?

Suggested Readings

Alkin, M. C. (Ed.). (2004). *Evaluation roots*. Thousand Oaks, CA: Sage.

House, E. R., & Howe, K. R. (1999). *Values in evaluation and social research.* Thousand Oaks, CA: Sage.

Reichardt, C. S., & Rallis, S. F. (Eds.). (1994). *The qualitative-quantitative debate: New perspectives.* New Directions for Program Evaluation, No. 61. San Francisco: Jossey-Bass.

Sechrest, L., & Figueredo, A. J. (1993). Program evaluation. *Annual Review of Psychology, 44,* 645–674.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1995). *Foundations of program evaluation*. Newbury Park, CA: Sage.

Stufflebeam, D. L. (2001). *Evaluation models.* New Directions for Evaluation, No. 89. San Francisco: Jossey-Bass.

5

First Approaches: Expertise and Consumer-Oriented Approaches

Orienting Questions

1. What are the arguments for and against using professional judgment as the means for evaluating programs?
2. What are the different types of expertise-oriented approaches? How are they alike and how do they differ?
3. Why is accreditation of institutions of higher education controversial today? How do these controversies reflect the controversies that frequently arise in many evaluations?
4. How is the consumer-oriented evaluation approach like the expertise-oriented approach? How is it different?
5. How do these approaches influence the practice of evaluation today?

Everyone evaluates. As we discussed in Chapter 1, we all form opinions or make judgments about the quality of things we encounter. Such evaluations include everything from the meal we just finished eating or the movie or concert we saw last week to more serious endeavors—the program to help students at risk of dropping out at our high school or the parent contact program for parents new to our school. Our focus here is not on our individual judgments of something, but on evaluations that are more formal, structured, and public. We connect these personal evaluations with the more formal ones here, though, because the earliest evaluation approaches were concerned, almost exclusively, with judging the quality of something. Those judgments were often derived by

a group of individuals coming together to consider their criteria and the program or product to be judged.

The first modern-day approaches to evaluation were expertise-oriented and consumer-oriented evaluations. These approaches continue to be used today, though not so widely in the professional evaluation field. However, they have influenced the ways we think of evaluation and its purposes and methods. We will review each briefly, with a focus on the most widely used current method—accreditation—to illustrate the key principles of these approaches and how they affected, and continue to affect, evaluation practices.

The Expertise-Oriented Approach

The expertise-oriented approach to evaluation is probably the oldest type of formal, public evaluation and, as its name implies, it relies primarily on professional expertise to judge the quality of an institution, program, product, or activity. For example, the merits of a leadership training program for school principals could be assessed by experts from various fields including leadership, educational administration, and training who would observe the program in action, examine its materials and underlying theory, perhaps interview some trainers and participants, or, in other ways, glean sufficient information to render a considered judgment about its value.

In another case, the quality of a hospital could be assessed by looking at its special programs, its operating facilities, its emergency room operations, its in-patient operations, its pharmacy, and so on, by experts in medicine, health services, and hospital administration. They could examine facilities and equipment/supplies of the hospital, its operational procedures on paper and in action, data on the frequency and outcomes of different procedures, the qualifications of its personnel, patient records, and other aspects of the hospital to determine whether it is meeting appropriate professional standards.

Although professional judgments are involved to some degree in all evaluation approaches, this one is decidedly different from others because of its direct, open reliance on professional expertise as the primary evaluation strategy. Such expertise may be provided by an evaluator or by subject-matter experts, depending on who might offer most in the substance or procedures being evaluated. Usually one person will not own all of the requisite knowledge needed to adequately evaluate the program, institution, or agency. A team of experts who complement each other are much more likely to produce a sound evaluation.

Several specific evaluation processes are variants of this approach, including doctoral examinations administered by a committee, proposal review panels, site visits and conclusions drawn by professional accreditation associations, reviews of institutions or individuals by state licensing agencies, reviews of staff performance for decisions concerning promotion or tenure, peer reviews of articles submitted to professional journals, site visits of educational programs conducted at the behest of the program's sponsor, reviews and recommendations by prestigious blue-ribbon

TABLE 5.1 *Some Features of Four Types of Expertise-Oriented Evaluation Approaches*

Type of Expertise-Oriented Evaluation Approach	*Existing Structure*	*Published Standards*	*Specified Schedule*	*Opinions of Multiple Experts*	*Status Affected by Results*
Formal review system	Yes	Yes	Yes	Yes	Usually
Informal review system	Yes	Rarely	Sometimes	Yes	Usually
Ad hoc panel review	No	No	No	Yes	Sometimes
Ad hoc individual review	No	No	No	No	Sometimes

panels, and even the critique offered by the ubiquitous expert who serves in a watchdog role.

To impose some order on the variety of expertise-oriented evaluation activities, we have organized and will discuss these manifestations in four categories: (1) formal professional review systems, (2) informal professional review systems, (3) ad hoc panel reviews, and (4) ad hoc individual reviews. Differences in these categories are shown in Table 5.1, along the following dimensions:

1. Is there an existing structure for conducting the review?
2. Are published or explicit standards used as part of the review?
3. Are reviews scheduled at specified intervals?
4. Does the review include opinions of multiple experts?
5. Do results of the review have an impact on the status of whatever is being evaluated?

Developers of the Expertise-Oriented Evaluation Approach and Their Contributions

It is difficult to pinpoint the origins of this approach, since it has been with us for a very long time. It was formally used in education in the 1800s, when schools began to standardize college entrance requirements. Informally, it has been in use since the first time an individual to whom expertise was publicly accorded rendered a judgment about the quality of some endeavor—and history is mute on when that occurred. Several movements and individuals have given impetus to the various types of expertise-oriented evaluations.

Elliot Eisner, an early evaluator discussed later in this chapter, stressed the role of connoisseurship and criticism in evaluation, roles that required expertise in the subject matter to be evaluated. James Madison and Alexander Hamilton took on the role of "expert evaluators" in discussing and elaborating on the meaning and merits of the newly proposed Constitution in *The Federalist Papers*. (They were experts because they were both present and active at the Constitutional Convention that drafted the document. As such, they were also

internal evaluators!) Their writings were influential at the time and are still used by jurists in the U.S. courts to interpret the meanings of the Constitution, illustrating the important actions that can come from reasoned judgments by experts about a product. Accreditation of institutions of higher education is the primary present-day application of expertise-oriented evaluations. The New England Association of Schools and Colleges, which granted the first accreditation and continues accreditations for colleges and universities in New England today, began in 1885 when a group of headmasters of preparatory secondary schools began meeting with presidents of colleges in New England to discuss what graduates should know to be prepared for college. Thus, more than 100 years ago, school and college leaders were talking about ways to align their curricula!

Formal Professional Review Systems: Accreditation

Historical Foundations. To many, the most familiar formal professional review system is that of accreditation, the process whereby an organization grants approval of institutions such as schools, universities, and hospitals. Beginning in the late 1800s, regional accreditation agencies in the United States gradually supplanted the borrowed western European system of school inspections. These agencies became a potent force in accrediting institutions of higher education during the 1930s. Education was not alone in institutionalizing accreditation processes to determine and regulate the quality of its institutions. Parallel efforts were under way in other professions, including medicine and law, as concern over quality led to wide-scale acceptance of professionals judging the efforts of those educating fellow professionals. Perhaps the most memorable example is Flexner's (1910) examination of medical schools in the United States and Canada in the early 1900s, which led to the closing of numerous schools he cited as inferior. As Floden (1983) has noted, Flexner's study was not accreditation in the strict sense, because medical schools did not participate voluntarily, but it certainly qualified as accreditation in the broader sense: a classic example of private judgment evaluating educational institutions.

Flexner's approach differed from most contemporary accreditation efforts in two other significant ways. First, Flexner was not a member of the profession whose efforts he presumed to judge. An educator with no pretense of medical expertise, Flexner nonetheless ventured to judge the quality of medical training in two nations. He argued that common sense was perhaps the most relevant form of expertise:

> Time and time again it has been shown that an unfettered lay mind, is . . . best suited to undertake a general survey. . . . The expert has his place, to be sure; but if I were asked to suggest the most promising way to study legal education, I should seek a layman, not a professor of law; or for the sound way to investigate teacher training, the last person I should think of employing would be a professor of education. (Flexner, 1960, p. 71)

It should be noted that Flexner's point was only partially supported by his own study. Although he was a layman in terms of medicine, he was an educator, and his judgments were directed at medical education rather than the practice of medicine, so even here appropriate expertise seemed to be applied.

Second, Flexner made no attempt to claim empirical support for the criteria or process he employed, because he insisted that the standards he used were the "obvious" indicators of school quality and needed no such support. His methods of collecting information and reaching judgments were simple and straightforward: "A stroll through the laboratories disclosed the presence or absence of apparatus, museum specimens, library, and students; and a whiff told the inside story regarding the manner in which anatomy was cultivated" (p. 79).

Third, Flexner dispensed with the professional niceties and courteous criticisms that often occur in even the negative findings of today's accreditation processes. Excerpts of his report of one school included scathing indictments such as this: "Its so-called equipment is dirty and disorderly beyond description. Its outfit in anatomy consists of a small box of bones and the dried-up, filthy fragments of a single cadaver. A cold and rusty incubator, a single microscope, . . . and no access to the County Hospital. The school is a disgrace to the state whose laws permit its existence" (Flexner, 1910, p. 190).

Although an excellent example of expertise-oriented evaluation (if expertise as an educator, not a physician, is the touchstone), Flexner's approach is much like that of contemporary evaluators who see judgment as the sine qua non of evaluation and who see many of the criteria as obvious extensions of logic and common sense (e.g., Scriven, 1973).

Accreditation in Higher Education Today. Accreditation in the United States and in many other countries today meets our criteria for an expertise-oriented, formal review system. The systems make use of an existing structure (generally an independent regional or national accreditation organization in the United States or governmental agencies in other countries), standards published by the organization responsible for accreditation, a specified schedule (for example, reviews of institutions every 2, 5, or 10 years), and opinions of multiple experts, and the status of the institution, department, college, or school is affected by the results. Accreditation is an excellent example of expertise-oriented evaluation because it uses people with expertise in the subject matter of the program or institution to form a judgment regarding the quality of the entity to be evaluated. The accreditation of an institution or program provides consumers and other stakeholders with some indication of the quality of the institution, as judged by experts in the field, and may facilitate summative decisions. For example, many students use an institution's or program's accreditation status to aid their decisions about whether to apply to or attend an institution or program. Further, the feedback the accreditation process provides to the institution can be used for program and institutional improvement and decision making. Thus, the accreditation process serves a formative purpose as well.

Accreditation in the United States is most common for institutions of higher education.[1] We will spend a little time describing this process because it has recently become quite political and controversial, and even for those readers not involved in accreditation, the arguments illustrate the types of political issues and choices that often arise in any evaluation. These include disagreements over the purpose of the evaluation (formative or summative); the neutrality and independence of the experts or evaluators; the criteria to be used to judge the product and, thus, the data to be collected or reviewed; and the transparency of the process (what should be available to the public or other stakeholders outside the organization). These controversies have emerged as the U.S. Department of Education, which has a stake in accreditation through provision of student loans to accredited institutions, has begun to take issue with the accreditation practices of the independent regional accrediting bodies that have traditionally reviewed colleges and universities for accreditation.

As noted earlier, in many countries, including Germany, the Netherlands, India, and the countries of the United Kingdom, institutions of higher education are required by law to be accredited. Government agencies, generally through a ministry or department of education, conduct the accreditation process. In some countries, such as Canada, there is no accreditation process for higher education, partly because most institutions of higher education are run by the provincial governments and that governance is considered to provide sufficient oversight. In the United States, accreditation evolved in a way that very much mirrors U.S. citizens' distrust of government. With a desire to minimize government's role nonprofit or voluntary associations carry out the accreditation tasks often fulfilled by government agencies in other countries.

As noted earlier, the New England Association of Schools and Colleges was the first accreditation organization in the United States. Originally established as a mechanism for dialogue between administrators of secondary schools and leaders of colleges in the region in 1885, it eventually evolved into the accrediting association for colleges and institutions in the region (Brittingham, 2009). Other regional associations followed, with each taking responsibility for accrediting institutions of higher education in their region. Today, there are six regional accrediting organizations in the United States, each pursuing similar activities within their region.[2] These

[1]Secondary institutions and school districts are occasionally accredited as well. Some states, for example, are moving to review school districts for accreditation and associations such as AdvancED have been formed out of the North Central and Southern accrediting associations for higher education to focus on accrediting K–12 schools. Further, many private schools are accredited. Our focus is on accreditation in higher education because it has been established for the longest period and its traditions, therefore, illustrate much about expertise-oriented evaluation and its controversies.

[2]The major regional accrediting associations in the United States are the Middle States Association of Colleges and Schools, the New England Association of Schools and Colleges, the North Central Association of Colleges and Schools, the Northwest Association of Accredited Schools, the Southern Association of Colleges and Schools, and the Western Association of Schools and Colleges. Although other accrediting organizations exist (for example, for religious institutions), these regional accrediting associations are considered the primary accrediting bodies in the United States.

associations focus primarily on accrediting institutions of higher education, though often they are also involved in accrediting K–12 schools. Finally, there are many accrediting associations that review programs in particular disciplines rather than entire institutions. For example, the American Bar Association accredits law schools, the Association of American Medical Colleges accredits medical schools, and the National Council for Accreditation of Teacher Education (NCATE) accredits teacher education programs, with the Teacher Education Accreditation Council (TEAC) emerging as a recent competitor to NCATE.

Accreditation of institutions of higher education by the six regional associations has followed a similar plan and approach, the mission-based approach, since the 1950s. With the mission-based approach, accreditors focus on the extent to which the institution is pursuing and achieving its stated mission. Although each association also has standards for higher education that it uses in the evaluation, the mission-based approach reflects the philosophy of the associations in its evaluations. Barbara Brittingham describes the mission-based approach and the accreditation process in the United States as "unusually focused on the future" to help the institution improve (2009, p. 18).

The Process of Accreditation. In the first stage of accreditation, the institution prepares a self-study report describing its mission and its progress toward that mission, as well as how the institution meets the standards of the accrediting body. The second major stage is the core of the expertise-oriented approach: a team of peers, faculty, and administrators from other institutions in the region receives the report and conducts a site visit during which they interview faculty, administrators, staff, and students; review institutional records on admissions, course curricula, student satisfaction and outcomes; observe facilities and classrooms, and so forth. Based on their review of the report and their experiences during the site visit, the team, usually three or four experts, writes a report expressing their views regarding the institution, their recommendations concerning its accreditation status, and their suggestions for improvement. The site visit report is then reviewed by a standing commission at the accrediting association, which may amend the conclusions. The commission then presents the final conclusions to the institution.

The process is expertise-oriented in several ways: (a) the association has expertise concerning standards for higher education, the state and status of other institutions, and the practice of accreditation and review; (b) the faculty and administrators who form the site team have expertise in participating in the governance of their own universities and others where they have been employed and receive some training from the association to serve as site reviewers. Therefore, the expertise of the site visit team and the association allows those involved to make use of the standards of the association, their review of the report, and their site visit to form a final judgment of the quality of the institution. This process is a common one followed not only by the regional accrediting organizations but also by the organizations that accredit programs in individual disciplines in higher education and by organizations that accredit other educational institutions, including school districts, private schools, charter schools, secondary schools, vocational schools, and religious schools.

Accreditation Controversies: Accreditation Politicized. So what can be controversial here? As one author defending the system notes, "Who better, one might ask, to evaluate the quality of a college or university than those who work in the field?" (O'Brien, 2009, p. 2). O'Brien argues that the evaluation and the relationship between the accrediting organizations and the institution should not be adversarial, noting, "The evaluators are not inspectors coming in with their white gloves" (O'Brien, 2009, p. 2). But the history of the controversy traces back to the GI Bill passed by Congress after World War II to provide financial assistance to returning soldiers to attend colleges and universities. The government wanted to ensure that the financial assistance went for worthwhile post secondary educational activities, but did not want to get directly into the business of examining colleges and universities for quality. So, it decided to rely on the independent regional accrediting associations, which were already reviewing colleges and universities, to determine the institutions students could receive financial aid to attend. Today, with increasing costs of higher education and more and more students attending colleges and universities, U.S. loans to students are big business. The government continues to rely on regional accrediting associations to identify the institutions of higher education that are eligible for aid, but has an increasing stake in the quality of those processes given the large amounts of money distributed in student loans and other forms of aid. In addition, the institutions themselves have a large stake in the process, because many students would not attend an institution that is not accredited, for quality and financial aid reasons.

Through the Higher Education Act, originally passed in 1965, the U.S. government influences higher education in many areas, from student loans to access. In recent years, many in the U.S. Department of Education have become concerned that accreditations are not sufficiently rigorous in weeding out schools that are performing poorly. Even proponents of the system note that current regional accreditation in the United States carries a "light touch" compared with government evaluations of higher education conducted in other countries (Brittingham, 2009, p. 18).

In 2005, the U.S. Department of Education appointed the Commission on the Future of Higher Education to study four issues critical to higher education, one of which was accountability. In "The Need for Accreditation Reform," a paper prepared for that report, Robert Dickeson called the current U.S. system of accreditation, "a crazy-quilt of activities, processes, and structures that is fragmented, arcane, more historical than logical, and has outlived its usefulness. More important, it is not meeting the expectations required for the future" (2006, p. 1). He concluded that "any serious analysis of accreditation as it is currently practiced results in the unmistakable conclusion that institutional purposes, rather than public purposes, predominate" (Dickeson, 2006, p. 3). He recommended that Congress create a National Accreditation Foundation to accredit institutions of higher education. The final report of the Commission, called the Spellings Commission for then Secretary of Education Margaret Spellings, was quite critical of current accreditation processes (U.S. Department of Education, 2006, http://www2.ed.gov/about/bdscomm/list/hiedfuture/reports/final-report.pdf). The report inspired much controversy and discussion in the higher education community, with

organizations such as Phi Beta Kappa and the Association of American Colleges and Universities issuing statements both of support and concern regarding the report. The final 2008 amendment of the Higher Education Act ultimately chose to ignore some of these recommendations, but the concerns raised by the Commission will continue (O'Brien, 2009) and, for our purposes, reflect some of the political concerns raised about evaluation today and, in particular, about expertise-oriented evaluation.

The regional accrediting associations see their purpose in evaluating institutions of higher education as primarily formative, helping these institutions improve. They see these goals as the best way to serve institutions, their students, and the public. By helping colleges and universities to improve and better achieve their stated mission, the accrediting associations believe they are helping students to receive a better education. In contrast, the U.S. Department of Education's emphasis is summative. It is concerned with maintaining the U.S. position in higher education in the world and in providing educated and skilled graduates for the economy of the twenty-first century. The Department and other critics see the purpose of accreditation as providing parents, students, and other consumers with information to help them decide which institutions they should attend and where they should spend their tuition dollars. In other words, accreditation should help these consumers make summative decisions about which institutions to choose. Further, accreditation should help make summative decisions about which institutions should continue. One critic notes that in the 60 years since the GI Bill was passed, "a mere handful of schools have been shut down and those largely for financial reasons . . . Meanwhile, on the accreditors' watch, the quality of higher education is slipping" (Neal, 2008, p. 26). So, the accrediting associations have developed a process that is most useful for formative evaluation when critics see the primary purpose as summative.

Increasing Emphasis on Outcomes. Another area of disagreement concerns the factors that should be considered in accreditation. Today, the emphasis in education, and in much of evaluation around the world, is on outcomes and impacts. (See Chapter 2.) The Spellings Commission report notes the following:

> Too many decisions about higher education—from those made by policymakers to those made by students and families—rely heavily on reputation and rankings derived to a large extent from inputs such as financial resources rather than outcomes. Better data about real performance and lifelong learning ability is absolutely essential if we are to meet national needs and improve institutional performance. (U.S. Department of Education, 2006, p. 14)

Just as K–12 education has moved to measuring student learning by focusing almost entirely on the extent to which state standards are achieved, the Spellings Commission would like evaluations of institutions of higher education to rely much more heavily on measures of student outcomes.[3] Although regional accrediting

[3]One difference between standards for K–12 education and those for higher education is that the standards for higher education would be national ones, not developed at the state level as K-12 standards are.

associations have begun to require institutions to provide measures of student outcomes and, for accreditations of professional programs, evidence concerning passage of licensing exams or job placements, the regional accreditation process also emphasizes the importance of input and process variables. Input variables include factors such as the quality of faculty, library holdings, IT capacity, classroom space and facilities, student admissions processes and decisions, and other elements that create the academic environment of the institution. Process variables articulated in standards, reviewed in self-reports, and examined by site visit teams include curricula, course requirements, and teaching quality; assistance to students through tutoring, advising, and other mechanisms; faculty-student interactions; internships; and other elements of the learning process. Regional accrediting associations also consider multiple outcomes, including graduation and drop-out rates, time to graduation, knowledge and skills of graduates, and job placements. Accrediting associations argue that they must examine the entire process of higher education to make a valid judgment of the quality of the institution and to provide advice for improvement. Examining only student outcomes does not give the experts in the accreditation process sufficient information to make useful recommendations for how to change the institution, and its inputs and processes, to achieve better outcomes (Murray, 2009).

Neutrality, Transparency, and Purpose in Accreditation. Other criticisms of the current approach concern reviewers' neutrality or objectivity and the transparency of the process. Evaluations are expected to be based on independent judgments. Such independence is intended to lead to more objective, and hence more valid, judgments of quality. Generally speaking, expertise-oriented evaluators should not be closely affiliated with the institution or product they are judging. For example, we are suspicious of an expert's endorsement of a product when we know the expert has a financial relationship with the product's manufacturer. Consider, for example, current discussions of the objectivity of medical research on the effectiveness of a drug when the research is funded by the pharmaceutical company that developed the drug. But accreditation processes make use of peer reviewers who are faculty and administrators from higher education institutions in the region. Accrediting organizations argue that these experts are in the best position to make the judgments and provide the advice institutions need, because they know what can be accomplished in the environment of such an institution—and how to accomplish it. They have worked in it themselves. Critics, however, are concerned that the closeness of the experts to those being judged and possible competition between institutions or departments present serious conflicts of interest that can lead to biased judgments. Judgments as blunt as Flexner's evaluations of medical schools would not see the light of day, at least in written reports.

Concerns over objectivity are heightened by the lack of transparency in the process. The U.S. Department of Education would like data and reports to be far more open, meaning that they would be available to parents, students, and the public and would contain content that is readily understood by nonexperts. For example, the Spellings Commission advocated tables presenting data on the

knowledge and skills of graduates and other outcome measures for various colleges and universities. These tables would be available for the public to use in judging the quality of institutions, and for other colleges to use as benchmarks (U.S. Department of Education, 2006). Accreditors rely on the thick descriptions contained in self-study reports and the accreditation report. Defenders of the current system agree that the system relies heavily on confidentiality but argue that this confidentiality is one of the reasons for its success. Because of it, "institutions can be candid in their self-studies, and teams can be honest in their assessments" (O'Brien, 2009, p. 2). If reports were made public, those writing the self-report would be reluctant to discuss real problems, and accreditation teams would edit their wording for public consumption. Neither would facilitate learning about problems and making recommendations for change.

Thus, accreditation is changing and is controversial. Like many evaluations in recent years, the accreditation of colleges and universities in the United States has moved to an increasing use of mixed methods and a greater focus on outcomes. Controversies concern the purpose of these expertise-oriented evaluations, the stakeholders they serve, the measures that should take priority, the neutrality and objectivity of the judgments of quality, the transparency of the process, and the availability of results to different stakeholders. Regional accrediting associations, which for many years had no competition, are being seriously challenged, not only by the federal government, but also by popular ratings of colleges and universities such as those published by *U.S. News and World Report*. As a result, accrediting associations are adapting and changing, but, with all their problems, they still remain a useful example of a formal review system using the expertise-oriented evaluation approach.

Other Formal Review Systems. There are numerous examples of other formal review systems, particularly in education. For many years, the National Council for Accreditation of Teacher Education (NCATE) has been the primary body to accredit teacher education programs. In 2000, this organization began focusing more on outcomes of such programs by examining knowledge and skills of graduates of the program, scores on licensure tests, and evidence that graduates are able to transfer their knowledge and skills to the classroom. The Teacher Education Accreditation Council (TEAC) has emerged as a competitor to NCATE, but with a similar focus on outcomes (Gitomar, 2007; Murray, 2009).

Some states are beginning to develop systems to review and accredit school districts within their state. For example, the Colorado Department of Education began accrediting districts in 1999 and revised the procedures substantially in 2008. The focus is very much on student outcomes and growth, but includes standards concerning "safe and civil learning environments," and budget and financial management. Reviewers conclude the process by assigning a district a rating at one of six different levels, from accreditation with distinction to probation and nonaccreditation. Like other formal review systems, the Colorado accreditation process for school districts includes published standards, specified schedules for review (annual for districts with lower ratings, 2 to 3 years for districts at higher levels of

accreditation), site visits by a team of external experts, and the districts' status being affected by the results (http://www.cde.state.co.us/index_accredit.htm).

Informal Review Systems

Many professional review systems have a structure and a set of procedural guidelines, and use multiple reviewers. Yet some lack the published standards or specified review schedule of a formal review system.

A graduate student's supervisory committee for dissertations, theses, or capstone projects is typically composed of experts in the student's chosen field and is an example of an informal system within expertise-oriented evaluation. Structures within the university, and/or faculty policies, exist for regulating such professional reviews of competence, but the committee members typically determine the standards for judging each student's performance. Fitzpatrick and Miller-Stevens (2009) have described the development and use of a rubric to assess students' performance on capstone projects to complete a master's program in public administration. But, typically, such criteria do not exist. Instead, the multiple experts on the committee make judgments of the student's performance, often without discussing their criteria explicitly. And, of course, the status of students is affected by the results.

The systems established for peer reviews of manuscripts submitted to professional periodicals might also be considered examples of informal review systems, though journals' procedures vary. Many journals do use multiple reviewers chosen for their expertise in the content of the manuscript. Unlike site visit teams for accreditation or members of a dissertation committee, reviewers do not behave as a team, discussing their reviews and attempting to reach consensus. Instead, a structure exists in the form of an editor or associate editor who selects reviewers, provides a timeframe for their reviews, and makes a final judgment about the manuscript based on the individual reviewers' comments. However, the schedule, like that for a graduate student's defense of a dissertation or thesis, is based on the receipt of manuscripts, although reviewers are given a specified time period in which to conduct the review. Many journals, but not all, provide reviewers with some general standards. Of course, the status of the manuscript—whether it is published, revised, or rejected—is affected by the review process.

Ad Hoc Panel Reviews

Unlike the ongoing formal and informal review systems discussed previously, many professional reviews by expert panels occur only at irregular intervals when circumstances demand. Generally, these reviews are related to no institutionalized structure for evaluation and use no predetermined standards. Such professional reviews are usually one-shot evaluations prompted by a particular, time-bound need for evaluative information. Of course, a particular agency may, over time, commission many ad hoc panel reviews to perform similar functions without their collectively being viewed as an institutionalized review system.

Panels to Develop Standards. Common examples of ad hoc review panels include panels organized in each state in the United States to develop or revise educational standards for a state or school district, funding agencies to judge proposals and make recommendations for funding, and blue-ribbon panels appointed to address particular issues. These ad hoc panel reviews have no routine schedule, but are organized by an agency or organization to receive input from experts on a particular issue. Thus, each of the 50 states has established standards that reflect that state's expectations regarding what students will know in different subjects at different grades.[4] There is considerable variation across the states in their standards, but the standards for each state were originally developed by a panel of experts. These experts typically consist of teachers, educational administrators, policymakers, and experts in the content area. The composition of the committee is intended to include experts with knowledge of the subject matter for which standards are being set and knowledge of the target population. Some sophisticated methods have been developed for the related task of expert committees identifying the cut scores, or scores that divide various test takers into groups based on their performance (Kane, 1995). (See Girard & Impara [2005] for a case study of the cut setting process by an expert panel in a public school district.)

Funding Agency Review Panels. In the United States, most federal government agencies make use of funding panels—panels of experts in the research area to be funded—to read proposals, discuss them, and make recommendations. Generally, the funding agency has developed criteria for the reviewers and, often, members of the team meet in Washington, DC, or other locations to discuss their reactions and attempt to reach some consensus. But the standards for funding vary from discipline to discipline and with the particular funding emphasis. Nevertheless, in the model of expertise-oriented evaluation, experts are coming together to make a judgment about something. Some funding organizations compose committees whose members have different areas of expertise. Thus, committees to review proposals in education can consist of a mix of educational administrators or policymakers, teachers, and researchers. Likewise, committees that review proposals for community development or action can include research experts in the field as well as community members serving as experts on the particular community and its needs.

Blue-Ribbon Panels. Blue-ribbon panels are typically appointed by a high-level government official and are intended to provide advice, not on funding, but on how government should address a particular issue. The Commission on the Future of Higher Education, which was discussed earlier in this chapter, was appointed by the U.S. Department of Education in 2005, at a time when the

[4]These actions are somewhat in response to the federal legislation commonly known as No Child Left Behind, but many states had developed standards prior to the legislation.

government was concerned with the long-term status of higher education in the United States and needed input from experts in the area. Members of such panels are appointed because of their experience and expertise in the field being studied. They typically are charged with reviewing a particular situation, documenting their observations, and making recommendations for action. Given the visibility of such panels, the acknowledged expertise of panel members is important if the panel's findings are to be considered credible. At the local level, where ad hoc review panels are frequently used as an evaluative strategy for many endeavors ranging from economic development and environmental policies to school governance, expertise of panel members is no less an issue, even though the reviewers may be of local or regional repute rather than national renown. Although recommendations of ad hoc panels of experts may have major impact, they might also be ignored, since there is often no formalized body charged with following up on their advice.

Ad Hoc Individual Reviews

Another form of expertise-oriented evaluation is the individual, professional review of any entity by any individual selected for his or her expertise to judge the value of the entity and, in some cases, to make recommendations for change or improvement. Employment of a consultant to perform an individual review of some educational, social, or commercial program or activity is commonplace in many organizations.

Educational Connoisseurship and Criticism

In the previous section, we discussed applications of the expertise-oriented approach in which the experts are not necessarily evaluators. They are experts in something else—the content they are judging. Further, these applications are examples of the expertise-oriented approach, but they were formed and exist independent of the professional evaluation community. In other words, we can study these processes as examples of expertise-oriented evaluation approaches, but those in the evaluation community are generally not involved in establishing these activities or in conducting them, as is the case with the other approaches we will discuss. As noted, we have begun our discussion of approaches by focusing on the oldest evaluation approach, one used for centuries before formal program evaluation emerged, to make judgments about important issues.

But, the expertise-oriented approach has also been part of the discussion of evaluation theories. In the early days of evaluation, Elliot Eisner was a key figure in discussing what evaluation should be, and his writings provide the theoretical foundation for the expertise-oriented approach and connect it to the evaluation literature (Eisner, 1976, 1985, 1991a, 1991b, 2004). Alkin and Christie (2004), in their evaluation tree depicting the origins and theories of evaluation, place Eisner, along with Michael Scriven, at the base of the valuing branch because their emphasis was on the valuing role of evaluation—determining the value, the merit or

worth, of the thing being evaluated. Eisner drew from the arts to describe his approach to evaluation. His perspective was a useful counterpoint to the emphasis in the 1970s on social science methods and program objectives. We will briefly discuss his concepts of connoisseurship and criticism, the fundamentals of his evaluation approach. These concepts fall within the expertise-oriented approach, because they require expertise in identifying and judging critical components or elements of the thing being evaluated.

The roles of the theater critic, art critic, and literary critic are well known and, in the eyes of many, useful roles. Critics are not without their faults. We may disagree with their views, but their reviews are good examples of direct and efficient application of expertise to that which is judged. Their criticism prompts us to think about the object being evaluated in different ways, even if we continue to disagree with their judgment. That is one goal of a written review or criticism: To prompt us to think about elements of the object that we, as nonexperts, might not have considered. Eisner (1991a) proposes that experts, like critics of the arts, bring their expertise to bear in evaluating the quality of programs in their areas of proficiency. Eisner does not propose a scientific paradigm but rather an artistic one, which he sees as an important qualitative, humanistic, nonscientific supplement to more traditional inquiry methods. He argues that we need to see the thing being evaluated from multiple perspectives and that the emphasis on quantitative, reductionist methods fails to convey many important qualities of the whole. He notes that numbers play a role in educational evaluation, his area of interest, but also limit what we see:

> [W]e should be recognizing the constraints and affordances of any form of representation we elect to use. Just as a way of seeing is also a way of not seeing, a way of describing is also a way of not describing. The tools we employ for noticing have an enormous impact on what it is that we become aware of. If we want a replete, fulsome, generous, complex picture of a classroom, a teacher, or a student, we need approaches to that perception of such phenomena and, in addition, a form of presentation that will make those features vivid. (Eisner, 2004, p. 200)

The key elements of Eisner's approach are connoisseurship and criticism (Eisner, 1975, 1991b). Connoisseurship is the art of appreciation—not necessarily a liking or preference for that which is observed, but rather an ability to notice, "to recognize differences that are subtle but significant in a particular qualitative display" (Eisner, 2004, p. 200). The connoisseur has developed knowledge of the important qualities of the object and the ability to observe and notice them well and to study the relationships among them. The connoisseur, in Eisner's view, is aware of the complexities that exist in observing something in real-world settings and possesses refined perceptual capabilities that make the appreciation of such complexity possible. The connoisseur's perceptual acuity results largely from a knowledge of what to look for (advance organizers or critical guideposts) gained through extensive previous experience, education, and reflection on that experience.

The analogy of wine tasting is used by Eisner (1975) to show how one must have many experiences to be able to distinguish what is significant about a wine, using a set of techniques to discern qualities such as body, color, bite, bouquet, flavor, and aftertaste, to judge its overall quality. The connoisseur's refined palate and gustatory memory of other wines tasted is what enables him or her to distinguish subtle qualities lost on an ordinary drinker of wine and to render judgments rather than mere preferences. Connoisseurs exist in all realms of life, not solely the gustatory or artistic. Eisner describes a good coach as a connoisseur of the game who, when watching others at the sport, can recognize subtleties that those with less experience would miss: "We see it displayed in blazing glory in watching a first-rate basketball coach analyze the strengths of the opponents, their weaknesses, as well as the strengths and weaknesses of the team that he or she is coaching" (2004, p. 198).

Connoisseurship does not, however, require a public description or judgment of that which is perceived. The public description is the second part of the Eisner approach. "Criticism," Eisner states, "is the art of disclosing the qualities of events or objects that connoisseurship perceives" (1979a, p. 197), as when the wine connoisseur either returns the wine or leans back with satisfaction to declare it of acceptable, or better, quality. Or, more akin to public evaluation, criticism is when the wine critic writes a review of the wine. Evaluators are cast as critics whose connoisseurship enables them to give a public rendering of the quality and significance of that which is evaluated. Criticism is not a negative appraisal but rather an educational process intended to enable individuals to recognize qualities and characteristics that might otherwise have been unnoticed and unappreciated. Criticism, to be complete, requires description, interpretation, and evaluation of that which is observed. "Critics are people who talk in special ways about what they encounter. In educational settings, criticism is the public side of connoisseurship" (Eisner, 1975, p. 13). Program evaluation, then, becomes program criticism. The evaluator is the instrument, and the data collecting, analyzing, and judging are largely hidden within the evaluator's mind, analogous to the evaluative processes of art criticism or wine tasting. As a consequence, the expertise—training, experience, and credentials—of the evaluator is crucial, because the validity of the evaluation depends on the evaluator's perception. Yet different judgments from different critics are tolerable, and even desirable, since the purpose of criticism is to expand perceptions, not to consolidate all judgments into a single definitive statement.

Eisner's educational criticism focuses on four dimensions that should be portrayed in a criticism: description, development of themes, interpretation, and evaluation. The focus is on expert, and sometimes, detailed description of the factors that are important in judging the quality of the product or program. Obviously, the approach would not be the most direct for clearly establishing cause-and-effect relationships, but it can be useful in helping us to understand the nature of the intervention and the manner in which it leads to different outcomes. As Eisner recently stated, "Educational connoisseurship and educational criticism represent an effort to employ what the arts and humanities as partners with the social sciences have to offer in advancing our understanding of the process and effect of education. In an age of high-stakes testing, it is a perspective we badly need" (Eisner, 2004, p. 202).

Influences of the Expertise-Oriented Approach: Uses, Strengths, and Limitations

Expertise-oriented approaches, generally referred to by other names, are used extensively in the United States and other countries today. Accreditation efforts are changing and expanding. Governments continue to appoint expert commissions to study issues and make recommendations. Often, such commissions help to protect government leaders from the ire of citizens when government needs to address a controversial issue. For example, closing military bases in the United States has been a controversial issue, in spite of the fact that too many bases exist. Congress and the president have resorted to appointing commissions of experts to provide "objective, non-partisan, and independent reviews" of recommendations for major base closures (http:www.brac.gov, homepage). The process has been used five times since the first commission was appointed in 1988, most recently in 2005. Like many blue-ribbon panels, the commissions have included experts in a variety of areas related to the issue. The commissions conduct site visits, seek input from the public and other experts, review information, and make recommendations to the President. The recommendations take effect unless Congress rejects the proposal within 45 days. These commissions have been able to take important actions to improve the efficiency and effectiveness of the placement of military bases.

Collectively, expertise-oriented approaches to evaluation have emphasized the central role of expert judgment, experience, and human wisdom in the evaluative process and have focused attention on such important issues as whose standards (and what degree of transparency) should be used in rendering judgments about programs. Conversely, critics of this approach suggest that it may permit evaluators to make judgments that reflect little more than personal biases. Others have noted that the presumed expertise of the experts is a potential weakness. Those using or contracting for expertise-oriented evaluations should consider carefully the various areas of expertise required for their team of expert judges. Too often the team contains only content experts, people who know various elements of the subject matter to be judged, but may lack experts in the evaluation process itself. The articulation of standards, whether by the contracting organization or by the team of experts, is also important to clarify the criteria and methods used to make the judgments requested. Of course, as Elliot Eisner would argue, experts should look beyond the standards and use their connoisseurship to describe, interpret, and judge the dimensions they know to be important to the quality of the product. But, articulated standards help to introduce some consistency across experts and to facilitate useful discussions among the experts when disagreements do occur.

Eisner's writings influenced evaluators to think more about the nature of evaluation judgments and the role that experience and connoisseurship can play in helping them to notice important elements of the program or product to be evaluated. However, Eisner did not remain active in the evaluation field, and the approach was used infrequently, generally by his immediate students. Still, we continue to study his writings because of the influences he has had on evaluation practice today. Donmoyer (2005) notes that Eisner's contributions prompted

evaluators to consider different approaches to evaluation and the implications of each. Eisner also provided an important rationale for qualitative methods at a time when quantitative methods dominated the field. His work was useful in prompting us to consider what we notice in an object. Connoisseurs know the important elements of a particular thing and learn how to form educated opinions about those elements. The connoisseurship-criticism approach also has its critics. Following Eisner's initial proposals, House (1980) issued strong reservations, cautioning that the analogy of art criticism is not applicable to at least one aspect of evaluation:

> It is not unusual for an art critic to advance controversial views—the reader can choose to ignore them. In fact, the reader can choose to read only critics with whom he agrees. A public evaluation of a program cannot be so easily dismissed, however. Some justification—whether of the critic, the critic's principles, or the criticism—is necessary. The demands for fairness and justice are more rigorous in the evaluation of public programs. (p. 237)

However, more recently, Stake and Schwandt emphasize the importance to evaluation not only of measuring quality but also of conveying quality as it is experienced. Reminiscent of Eisner's recognition of connoisseurship, they observe that "we do not have good enough standards for recognizing an evaluator's practical knowledge that arises from a combination of observational skill, breadth of view, and control of bias" (2006, p. 409). They conclude that "as with connoisseurs and the best blue ribbon panels, some of the best examples of synthesizing values across diverse criteria are those that rely on the personal, practical judgment of fair and informed individuals" (2006, p. 409).

The Consumer-Oriented Evaluation Approach

Like the expertise-oriented approach, consumer-oriented evaluation has existed in the practice of individuals making decisions about what to purchase, or trade, for centuries. The approaches are similar in other ways: Their primary purpose is to judge the quality of something, to establish the value, the merit or worth, of a product, program, or policy. Although all evaluations are concerned with determining merit or worth, valuing is the key component of these two approaches.[5] Their principal audience is the public. Unlike approaches that will be discussed in other chapters in this section, evaluations relying on these approaches often do not have another audience—a foundation, manager, policymaker, or citizens' group—who has hired the evaluator to provide them with useful information to

[5]Other evaluation approaches focus on various types of use, such as stakeholder involvement or organizational change, and methodology, such as establishing causality or providing thick descriptions as the central component. These evaluations, too, may ultimately make a judgment of merit or worth, but that judgment, the valuing of the program or product, is not so central to the evaluation approach as it is in expertise-oriented or consumer-oriented evaluation. (See Alkin [2004], Shadish et al. [1991].)

make a decision or judgment. Instead, the audience for consumer-oriented and expertise-oriented approaches is a broader one—the purchasing or interested public—and is not directly known to the evaluator. Therefore, the evaluator is the major, often the only, decision maker in the study because he or she does not have other important, direct audiences to serve. But the consumer-oriented approach and the expertise-oriented approach differ dramatically in their methodologies, with the latter relying on the judgments of experts and the arts as a model. On the other hand, consumer-oriented evaluation relies on more transparent and quantitative methods, with the judgment typically being made by an evaluator, a person with expertise in judging things, but not with the particular content expertise of expertise-oriented or connoisseur evaluations.

Popular examples of consumer-oriented evaluations that the reader will know include *Consumer Reports* and the *U.S. News and World Repor*t ratings of colleges and universities, but examples exist around the world. *Which?* is a magazine and web site in the United Kingdom that serves a mission similar to that of the Consumers' Union, the sponsor of *Consumer Reports* and its web site, in the United States. Both organizations act as consumer advocates and test products to provide information to consumers on the effectiveness of various products.

The Developer of the Consumer-Oriented Evaluation Approach

Consumer-oriented evaluations first became important in educational evaluations in the mid to late 1960s as new educational products flooded the market with the influx of funds from the federal government for product development. Michael Scriven is the evaluator best known for prompting professional evaluators to think more carefully about consumer-oriented or product evaluations (1974b, 1991c). Scriven, of course, is known for many things in evaluation, and consumer-oriented or product-oriented evaluations represent only one of his contributions. His most important contributions include making evaluators aware of the meaning and importance of valuing in evaluation (Shadish et al. 1991; Alkin, 2004). He often uses examples of product evaluation in his writing to illustrate the nature of valuing and the process of deriving a value in evaluation. For many years, he considered *Consumer Reports* to be "an almost flawless paradigm" in product evaluation. However, he has expressed disappointment with their reluctance to discuss and improve their methodology and has recognized *PC Magazine* and *Software Digest* as developing more methodologically sound procedures (Scriven, 1991a, p. 281).

Scriven's approach to determining the value of a product, however, is quite different from Eisner's connoisseur approach. In fact, Scriven's critical view of Eisner's approach illustrates his own priorities. He states that evaluations using the connoisseurship model "may generate a valuable perspective, but it abandons much of the requirement of validity. In particular it is vulnerable to the fallacy of irrelevant expertise, because connoisseurs are at best a bad guide to merit for the novice—and are also affected by the swing of fashion's pendulum" (Scriven, 1991a, p. 92).

So, while Eisner's model rests on the noticing abilities attained by the connoisseur, Scriven's methods for product evaluation are not concerned with expertise in the content of the product, but with the evaluator's expertise in testing and judging key components of the product. Further, although Eisner emphasizes interpreting and evaluating the product, he believes that the value added of his approach is in the description—in helping others perceive, and experience, key elements they may have overlooked. Scriven's concern is in answering the question, "How good is this product?" To do so, he collects information to judge the product's performance and that of its competitors on explicit, critical criteria and works to remove subjectivity from the approach. Thus, he notes the procedures used by two consumer-oriented magazines he admires represent a "'pure testing' approach, that is, one which minimizes the amount of subjective judgment in a particular case" (Scriven, 1991a, p. 281).

Stake and Schwandt (2006), in a discussion of the importance of evaluators discerning quality, shed some light on the differences in Eisner's and Scriven's approaches. They identify two approaches to conceptualizing quality: quality as measured and quality as experienced. Quality as experienced is derived from practical knowledge and personal experience, and is significant, they argue, because it is the means by which many people determine quality. Eisner's connoisseurship model would appear to be an example of evaluation that builds on such quality, through the eyes and experience of a connoisseur. In contrast, quality as measured is illustrated in Scriven's logic of evaluation and his method for evaluating products. These include determining the important criteria to consider in evaluating the product, establishing standards for the criteria, examining or measuring the performance of the products and its competitors against the criteria using the standards, and synthesizing the results to determine the quality of the key product. Both views of quality have a role. We have discussed Eisner's approach. Let us now describe more of Scriven's model for judging the quality of a product.

Applying the Consumer-Oriented Approach

A key step in judging a product is determining the criteria to be used. In the consumer-oriented model, these criteria are explicit and are presumably ones valued by the consumer. Although Scriven writes about the possibility of conducting needs assessments to identify criteria, his needs assessments are not formal surveys of consumers to determine what they would like. Instead, his needs assessments focus on a "functional analysis" that he writes is "often a surrogate for needs assessments in the case of product evaluation" (Scriven, 1983, p. 235). By functional analysis, Scriven means becoming familiar with the product and considering what dimensions are important to its quality:

> Once one understands the nature of the evaluand, . . . one will often understand rather fully what it takes to be a better and a worse instance of that type of evaluand. Understanding what a watch is leads automatically to understanding what the dimensions of merit for one are—time-keeping, accuracy, legibility, sturdiness, etc. (1980, pp. 90–91)

Thus, his criteria are identified by studying the product to be evaluated, not by previous, extended experience with the product. Standards, developed next, are levels of the criteria to be used in the measurement and judgment process. They are often created or recognized when comparing the object of the evaluation with its competitors. Since the goal is to differentiate one product from another to inform the consumer about quality, standards might be relatively close together when competitors' performances on a criterion are similar. In contrast, standards might be quite far apart when competitors differ widely. Standards, of course, can be influenced by factors other than competitors, such as safety issues, regulatory requirements, and efficiency factors that provide common benchmarks.

Scriven's work in product evaluation focused on describing this process and, in part because identifying criteria can be difficult, in developing checklists of criteria for others to use in evaluating products. His product checklist published in 1974 reflects the potential breadth of criteria that he recommends using in evaluating educational products (Scriven, 1974b). This product checklist, which remains useful today, was the result of reviews commissioned by the federal government, focusing on educational products developed by federally sponsored research and development centers, and regional educational laboratories. It was used in the examination of more than 90 educational products, most of which underwent many revisions during the review. Scriven stressed that the items in this checklist were necessitata, not desiderata. They included the following:

1. *Need*: Number affected, social significance, absence of substitutes, multiplicative effects, evidence of need
2. *Market*: Dissemination plan, size, and importance of potential markets
3. *Performance—True field trials*: Evidence of effectiveness of final version with typical users, with typical aid, in typical settings, within a typical time frame
4. *Performance—True consumer*: Tests run with all relevant consumers, such as students, teachers, principals, school district staff, state and federal officials, Congress, and taxpayers
5. *Performance—Critical comparisons*: Comparative data provided on important competitors such as no-treatment groups, existing competitors, projected competitors, created competitors, and hypothesized competitors
6. *Performance—Long-term*: Evidence of effects reported at pertinent times, such as a week to a month after use of the product, a month to a year later, a year to a few years later, and over critical career stages
7. *Performance—Side effects*: Evidence of independent study or search for unintended outcomes during, immediately following, and over the long-term use of the product
8. *Performance—Process*: Evidence of product use provided to verify product descriptions, causal claims, and the morality of product use
9. *Performance—Causation*: Evidence of product effectiveness provided through randomized experimental study or through defensible quasi-experimental, expost facto, or correlational studies

10. *Performance—Statistical significance*: Statistical evidence of product effectiveness to make use of appropriate analysis techniques, significance levels, and interpretations
11. *Performance—Educational significance*: Educational significance demonstrated through independent judgments, expert judgments, judgments based on item analysis and raw scores of tests, side effects, long-term effects and comparative gains, and educationally sound use
12. *Cost-effectiveness*: A comprehensive cost analysis made, including expert judgment of costs, independent judgment of costs, and comparison to competitors' costs
13. *Extended Support*: Plans made for post-marketing data collection and improvement, in-service training, updating of aids, and study of new uses and user data

These criteria are comprehensive, addressing areas from need to process to outcomes to cost. Scriven also developed a checklist to use as a guide for evaluating program evaluations, the Key Evaluation Checklist (KEC) (Scriven, 1991c, 2007). It can be found at http://www.wmich.edu/evalctr/checklists/kec_feb07.pdf.

Other Applications of the Consumer-Oriented Approach

Product evaluation is also used by organizations and industries to evaluate products at many different stages. Successful high-technology companies such as Apple have watched and studied consumers' reactions to iPhones and Apple stores and used these data to make changes in their products, thus using consumer-oriented evaluations for formative purposes to revise their products. Amazon.com undertook a similar process with its electronic book, Kindle. Jonathan Morrell, an evaluator who has worked with industries to conduct many product evaluations, recently described the present-day use of product evaluations in industry. Although Scriven focused on product evaluations for summative, purchasing decisions by consumers, Morrell notes that most product evaluations in industries are formative in nature, as with the examples of Apple and Amazon.com. Evaluations take place through the product's life cycle from initial design and the production process to marketing and circulation. The stakeholders for the evaluation include not only the managers of the organization and the consumers, but others associated with the product process as well. Morrell gives the example of pilots as a stakeholder for airplanes. Their opinions on human factors issues are important in creating a product that will permit them to perform optimally in flying the plane (Morell, 2005).

Influences of the Consumer-Oriented Approach: Uses, Strengths, and Limitations

As mentioned previously, the consumer-oriented approach to evaluation has been used extensively by government agencies and independent consumer advocates to make information available on hundreds of products. One of the best known

examples in education today is the What Works Clearinghouse (WWC), begun in 2002 by the U.S. Department of Education's Institute for Education Sciences (IES). (See http://ies.ed.gov/ncee/wwc.) WWC is a source for consumer-oriented evaluation information on the outcomes of educational programs and products. Its intent, like the consumer-oriented approach reviewed here, is to help consumers—teachers, school psychologists, and educational administrators—make choices about which educational products to use.

WWC differs dramatically, however, from Scriven's more comprehensive evaluation process because its criteria for determining program success are confined to program outcomes, and its standards are concerned with research confidence in those outcomes. The stated mission of WWC is "to assess the strength of evidence regarding the effectiveness of the program."[6] Products studied using randomized control trials (RCTs) or regression discontinuity designs, which are viewed by IES as superior for establishing a causal link between the product or program and the outcome, receive the highest ratings. Studies using quasi-experimental designs may be endorsed with reservations. Scriven's checklists and writings argued for using several different criteria to reflect the elements of the product or program that were critical to successful performance. Although many of Scriven's criteria concerned outcomes or performance (see his criteria for judging educational products listed previously), his process emphasized a comprehensive appraisal of the product, including need, side effects, process, support for users, and cost, as well as several criteria concerning outcomes or performance. WWC's standards concern the extent to which the research establishes a causal effect, through preferred designs, between the program or product and the intended outcome. Although we bemoan the narrowing of the range of criteria and the standards to assess those criteria, WWC's efforts do prompt the potential user to consider the effectiveness of the program in achieving its outcomes and to provide a central location for accessing comparable information on educational programs and products. Educators are currently under much pressure to increase achievement, and products can mislead in their marketing. However, WWC's efforts to inform the consumer about the demonstrated success of programs and products is today's most successful application of the consumer-oriented approach in education in terms of visibility and number of users. Consumers can search the web site by area of interest, with topics including Early Childhood Education, Beginning Reading, Middle School Math, Dropout Prevention, and English Language Learners. Many products are judged to have insufficient research evidence for a causal relationship between the product and the outcome. The only information provided on these products is the designation "no studies meeting eligibility

[6]In an ironic combination of consumer-oriented and expertise-oriented approaches, a blue-ribbon panel was convened in 2008 to determine whether WWC's review process and reports were "scientifically valid" and "provide accurate information about the strength of evidence of meaningful effects in important educational outcomes." See http://ies.ed.gov/director/board/pdf/panelreport.pdf. Commenting that their charge was not to review the mission but to determine if the information was valid, the panel concluded that the information provided was valid.

standards." However, for products with studies meeting the eligibility standards, reports provide a brief description of the program or product, the research conducted on it, and a final judgment of its effectiveness at achieving the intended outcome.

Another prominent example of the consumer-oriented approach that illustrates the overlap between it and the expertise-oriented approach are the test reviews of the Buros Institute of Mental Measurements. The Institute was founded in 1938 and has been conducting well-respected reviews of educational and psychological tests since that time. It currently produces two series: *The Mental Measurements Yearbooks,* now in its 17th edition, and Test Reviews Online (see www.unl.edu/buros). The Institute is consumer oriented in that it is "dedicated to monitoring the quality of commercially-published tests . . . promoting appropriate test selection, use, and practice" (http://www.unl.edu/buros/bimm/html/catalog.html, paragraph 1). It is designed to provide consumers with information on the quality of tests used in education and psychology. Each test review provides a brief description of the test and a discussion of its development and technical features, including reliability and validity information, a commentary, a summary, and references. However, the reviews contain elements of the expertise-oriented approach because they are conducted by experts in psychometrics and, although the reviews make use of a prescribed format, the criteria and standards for reviewing each test and its competitors are not explicitly identified as would be done in Scriven's approach. The Institute encourages its reviewers to use *The Standards for Educational and Psychological Testing* (1999), developed jointly by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), as a guide, but the Institute's primary criterion for providing information on quality are in the selection of its expert reviewers.

Although the consumer-oriented evaluation approach continues to be used by magazines and web sites that review products, the approach is not one that continues to be discussed extensively in the professional evaluator literature. However, Scriven's writings on product evaluation in the 1970s, as well as Eisner's writings on connoisseurship and criticism, were important in influencing evaluation in its early stages to consider its role in valuing a program, policy, or product and in considering methods other than traditional social science research methods, for doing so. Each approach influenced evaluation practice today.

Major Concepts and Theories

1. The hallmark of the expertise-oriented evaluation approach is its direct reliance on professional judgment in the area of the program being evaluated.

2. Variations in the types of expertise-oriented evaluations include formal and informal review systems and ad hoc panels or individual reviews. These evaluations vary as to whether they are housed under an existing structure or organization, have

published standards that are used to evaluate the program or product, use a predetermined schedule for review, employ single or multiple experts, and directly affect the status of the program.

3. Accreditation systems in higher education, extending to K–12 schools, are a prominent example of the expertise-oriented evaluation approach in the United States and are currently in a process of discussion and change. Differences between the regional accrediting associations in the United States and the federal government concerning the purposes of these evaluations, the nature of the data collected or reviewed (outcomes, process, and inputs), the independence or neutrality of the expert evaluators, and the transparency of the process illustrate many of the controversies and political issues that can arise in expertise-oriented and other evaluations.

4. Elliot Eisner's educational connoisseurship and criticism model made evaluators more aware of the skills of an expert, or connoisseur, in noticing critical dimensions of a product or program and in using methods outside of traditional social science measurement, especially qualitative methods of observation and description, to provide a complete picture of the program or product.

5. The consumer-oriented evaluation approach differs from the expertise-oriented approach in that it does not rely on content experts, or connoisseurs of the product, but rather on experts in evaluation. The approach is also based more centrally on evaluation logic and quantitative methods.

6. Michael Scriven, who wrote extensively about such evaluations, described the key steps as identifying the important criteria for judging the product or program, developing standards to judge those criteria, collecting information or data, and synthesizing the information to make a final judgment that permits the consumer to compare the product with likely alternatives.

7. Both expertise-oriented and consumer-oriented approaches made evaluators aware of the importance of valuing in their work. It helped them recognize that the central task of evaluation is to make a judgment about the value of a program, product, or policy. The approaches advocate quite different methods for making that judgment and, therefore, each added separately to evaluators' consideration of qualitative methods and of criteria, standards, and checklists as potential methods for collecting data.

8. Both approaches continue to be used commonly by public, nonprofit, and private organizations and industries, but are not the subject of much writing in professional evaluation today. The absence of evaluation literature on the subject is unfortunate. We hope evaluators will return their attention to these approaches commonly used by others to bring evaluative ways of thinking to the application of the approaches today.

Discussion Questions

1. How do expertise-oriented and consumer-oriented evaluation approaches differ? How are they alike?

2. What do you see as the strengths of the expertise-oriented approaches? What are their drawbacks?

3. If a team of experts were reviewing your school or organization, what kinds of experts would you want on the team? What criteria would you want them to use to judge the quality of your organization?
4. Referring to question 3, who would you trust to make a better judgment—someone who is an expert in the content or subject matter of your organization or someone who knows evaluation theories, and methods for judging something? Justify your response.
5. Discuss the concept of a connoisseur. Are you a connoisseur at something? What is it? How does your experience with this thing help you to notice the important factors and be able to judge them better than a novice?
6. In consumer-oriented evaluation, what is the difference in criteria and standards?
7. How should one determine the criteria for evaluating a product? Should the focus be solely or primarily on outcomes? What should be the balance among the quality of inputs (staff, facilities, budget), process (the conduct of the program), and outputs or outcomes?

Application Exercises

1. What outside experts review your program or organization?
   a. If you work in an organization that is accredited, review the standards used for accreditation. Do you feel the standards get at the real quality issues of the program or organization? What other standards might you add?
   b. What are the areas of expertise of the evaluation team? Are they content experts, management experts, finance experts, evaluation experts, or experts in other areas? How do you judge the mix of expertise? Might you add others? How might others judge their independence or objectivity in judging your organization?
   c. If possible, interview those involved in the accreditation and learn more about the purposes of the accreditation (whether the emphasis is formative, summative, or something else) and about how it has been used.
2. Your high school is going to be visited by an outside accreditation team. What issues do you think they should attend to? What do you think they might miss in a short visit? What information do you think they should collect? What should they do while they're visiting? Do you think such a team could make a difference for your school? Why or why not?
3. Read a review of a restaurant, movie, or play that you have attended or seen. How does your opinion differ from the critic's? How do the critic's opinions influence your own? Does his or her expertise in the product (connoisseurship) or his ability to communicate it (criticism) prompt you to think about the product in different ways?
4. Look at an evaluation of an educational product of interest to you on What Works Clearinghouse at http://ies.ed.gov/ncee/wwc. Critique their presentation of information from an expertise-oriented and from a consumer-oriented approach. What information is helpful? Would other information be helpful to you in making a decision? If so, what? Does that information relate to a different criterion or standard you have? How does the information fit into the approaches reviewed in this chapter?

5. The product or program you are interested in is not reviewed by What Works Clearinghouse, so you are going to contact the publisher or developer of this product to learn more about it. What criteria are important to you? What standards might you use to judge those criteria? What will you ask the person who represents the company?

6. Examine a recent issue of *Consumer Reports* or a similar magazine or online publication that reviews products and critique their review of a particular product. Do you agree with their selection of the criteria to judge the product? Would you exclude any criteria? Include others? Are the standards they use to judge each product on the criteria explicit? Appropriate? How would you judge their data collection process, that is, their means for determining how each product performs on the criteria? As an expert, or perhaps a connoisseur of consumer-based evaluation, how would you judge their evaluation? How would you improve their process?

A Case Study

For this chapter, we recommend an interview with Gary Henry on the development of the Georgia school report card in *Evaluation in Action*, Chapter 7. Although our interviews do not contain any evaluations that explicitly use an expertise-oriented or consumer-oriented approach, this interview illustrates the development of a school report card to be used by consumers, parents, and citizens of Georgia. Some of Dr. Henry's work is concerned with identifying and developing the multiple criteria to be used on the report card, using research studies and input from surveys of the citizens of Georgia and the advisory council to the evaluation. He discusses this process of identifying criteria in his interview and the means for formatting the information in an accessible, easy-to-use manner, and then disseminating it widely. The journal source is Fitzpatrick, J. L., & Henry, G. (2000). The Georgia Council for School Performance and its performance monitoring system: A dialogue with Gary Henry. *American Journal of Evaluation, 21*, 105–117.

Suggested Readings

Eisner, E. W. (1991a). Taking a second look: Educational connoisseurship revisited. In M. W. McLaughlin and D. C. Philips (Eds.), *Evaluation and education: At quarter century, Ninetieth Yearbook of the National Society for the Study of Education, Part II*. Chicago: University of Chicago Press.

Eisner, E. W. (1991b). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice*. New York: Macmillan.

Floden, R. E. (1980). Flexner, accreditation, and evaluation. *Educational Evaluation and Policy Analysis, 20*, 35–46.

O'Brien, P. M. (Ed.). *Accreditation: Assuring and enhancing quality*. New Directions in Higher Education, No. 145, pp. 1–6. San Francisco: Jossey-Bass.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.

Scriven, M. (2007). *Key evaluation checklist*. http://www.wmich.edu/evalctr/checklists/kec_feb07.pdf

U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. Washington, DC. http://www.ed.gov/about/bdscomm/list/hiedfuture/reports/final-report.pdf

6

Program-Oriented Evaluation Approaches

Orienting Questions

1. What are the key concepts of the objectives-oriented evaluation approach and how has this approach influenced evaluation? How is this approach used today?
2. How are logic models and program theories used in evaluation?
3. How does theory-based evaluation differ from objectives-oriented evaluation? What are the central concepts of theory-based evaluation?
4. What are some of the strengths and limitations of the major program-oriented evaluation approaches?
5. What is "goal-free evaluation"? What does it teach us about conducting evaluation?

Today many approaches to evaluation begin their focus with learning more about some key features of the program to be evaluated. These features then serve to help the evaluator decide which questions should be addressed. The most prominent program-oriented approaches are the objectives-oriented approach and approaches that make use of logic models or program theory. In fact, theory-based evaluation is one of the most rapidly growing areas of evaluation (Weiss, 1995; Donaldson, 2007). Many government funding agencies and foundations require logic models, a variant of program theory, for program planning, evaluation, and research. Both logic models and program theory have evolved to help evaluators gain a better understanding of the rationale or reasoning behind the program's intended effects; this represents a great improvement over the more traditional objectives-oriented evaluation, which focused only on stated program outcomes.

In this chapter, we will cover the original program-oriented evaluation approach—objectives-oriented evaluation—which continues to be used frequently today. We will then describe theory-oriented approaches and their cousin, logic models, and their applications today in helping evaluators make the critical choices of what to evaluate.

The Objectives-Oriented Evaluation Approach

The distinguishing feature of an objectives-oriented evaluation approach is that the purposes of some activity are specified, and then the evaluation focuses on the extent to which those purposes, or objectives, are achieved. In many cases, programs already have specified objectives. In other cases, the evaluator may work with stakeholders to articulate the program objectives, sometimes called goals or standards. The key role for the evaluator in an objectives-oriented evaluation is to determine whether some or all of the program objectives are achieved and, if so, how well they are achieved. In education, the objectives may be concerned with the purposes of a single lesson or training program or the knowledge students should attain during an entire year. In public health programs, the objectives may concern the effects of prevention efforts, community health interventions, or patient education. Objectives in environmental programs might include such quantitative outcomes as reduction in air pollutants or outcomes that are more difficult to measure such as citizens' beliefs and behaviors about energy use. The information gained from an objectives-oriented evaluation could be used to determine whether to continue funding the program, change significant portions of it, or throw it out and consider other approaches.

Many people have contributed to the evolution and refinement of the objectives-oriented approach to evaluation since its inception in the 1930s, but the individual most credited with conceptualizing and popularizing the focus on objectives in education is Ralph W. Tyler (1942, 1950).

The Tylerian Evaluation Approach

Tyler had a tremendous influence on both evaluation and education., His work influenced the Elementary and Secondary Education Act (ESEA) of 1965, the first federal act to require evaluation of educational programs. At the end of his career, he chaired the committee that started the National Assessment of Educational Progress (NAEP), which in the United States today remains the only way to examine educational achievement across all 50 states because of the different standards in each state. In the 1920s and 1930s, while working closely with teachers and schools, Tyler began to formulate his views on education and evaluation. His writings and work foreshadowed today's concepts of continuous improvement and multiple means of assessment. He saw objectives as a way for teachers to

define what they wanted students to learn. By stating objectives in terms of what students should be able to do, Tyler believed that teachers could more effectively plan their curricula and lessons to achieve those objectives. Unlike later versions of behavioral objectives, however, Tyler believed that objectives should concern principles, not minute behaviors. He worked closely and cooperatively as an evaluator with teachers to make evaluation and education cooperative endeavors (Goodlad, 1979; Madaus, 2004; Madaus & Stufflebeam, 1989).

Tyler considered evaluation to be the process of determining the extent to which the objectives of a program are actually being achieved. His approach to evaluation followed these steps:

1. Establish broad goals or objectives.
2. Classify the goals or objectives.
3. Define objectives in behavioral terms.
4. Find situations in which achievement of objectives can be shown.
5. Develop or select measurement techniques.
6. Collect performance data.
7. Compare performance data with behaviorally stated objectives.

Discrepancies between performance and objectives would lead to modifications intended to correct the deficiency, and the evaluation cycle would be repeated.

Tyler's rationale was logical, scientifically acceptable, readily adoptable by evaluators (most of whose methodological training was very compatible with the pretest-posttest measurement of behaviors stressed by Tyler), and had great influence on subsequent evaluation theorists. Tyler advocated multiple measures of different types and considered many elements of a program during an evaluation. However, the objectives-oriented approaches that evolved from Tyler's work in the 1960s and 1970s and that continue to be used in some settings today focused on a basic formula: articulate program objectives; identify the means, typically tests, to measure them; administer the tests; analyze the data in reference to previously stated objectives; and determine program success.

This basic, objectives-oriented approach is largely discredited by professional evaluators today. However, many funding sources have not caught up with present-day evaluation approaches and require evaluations to make use of this traditional approach. Its strengths and limitations are discussed in the conclusion of the chapter.

Provus's Discrepancy Evaluation Model

Another approach to evaluation in the Tylerian tradition was developed by Malcolm Provus, who based his approach on his evaluation assignments in the Pittsburgh public schools (Provus, 1971, 1973). Provus viewed evaluation as a continuous information-management process designed to serve as "the watchdog of program management" and the "handmaiden of administration in the management of program development through sound decision making" (Provus, 1973, p. 186). Although his was, in some ways, a management-oriented

evaluation approach, the key characteristic of his proposals stemmed from the Tylerian tradition. Provus viewed evaluation as a process of (1) agreeing on standards (another term used in place of objectives),[1] (2) determining whether a discrepancy exists between the performance of some aspect of a program and the standards set for performance, and (3) using information about discrepancies to decide whether to improve, maintain, or terminate the program or some aspect of it. He called his approach, not surprisingly, the Discrepancy Evaluation Model (DEM).

Provus determined that, as a program is being developed, it goes through four developmental stages, to which he added a fifth, optional stage:

1. Definition
2. Installation
3. Process (interim products)
4. Product
5. Cost-benefit analysis (optional)

During the definition, or design, stage, the focus of work is on defining goals and processes or activities and delineating necessary resources and participants to carry out the activities and accomplish the goals. Provus considered programs to be dynamic systems involving inputs (antecedents), processes, and outputs (outcomes). Standards or expectations were established for each stage. These standards were the objectives on which all further evaluation work was based. The evaluator's job at the design stage is to see that a complete set of design specifications is produced and that they meet certain criteria: theoretical and structural soundness.

At the installation stage, the program design or definition is used as the standard against which to judge program operation. The evaluator performs a series of congruency tests to identify any discrepancies between expected and actual implementation of the program or activity. The intent is to make certain that the program has been installed as it has been designed. This is important because studies have found that staff vary as much in implementing a single program as they do in implementing several different ones. The degree to which program specifications are followed is best determined through firsthand observation. If discrepancies are found at this stage, Provus proposed several solutions to be considered: (a) changing the program definition to conform to the way in which the program is actually being delivered if the actual delivery seems more appropriate, (b) making adjustments in the delivery of the program to better conform to the program definition (through providing more resources or training),

[1]Although *standards* and *objectives* are not synonymous, they were used by Provus interchangeably. Stake (1970) also stated that "standards are another form of objective: those seen by outside authority figures who know little or nothing about the specific program being evaluated but whose advice is relevant to programs in many places" (p. 185).

or (c) terminating the activity if it appears that further development would be futile in achieving program goals.

During the process stage, evaluation focuses on gathering data on the progress of participants to determine whether their behaviors changed as expected. Provus used the term "enabling objective" to refer to those gains that participants should be making if longer-term program goals are to be reached. If certain enabling objectives are not achieved, the activities leading to those objectives are revised or redefined. The validity of the evaluation data would also be questioned. If the evaluator finds that enabling objectives are not being achieved, another option is to terminate the program if it appears that the discrepancy cannot be eliminated.

At the product stage, the purpose of evaluation is to determine whether the terminal objectives for the program have been achieved. Provus distinguished between immediate outcomes, or terminal objectives, and long-term outcomes, or ultimate objectives. He encouraged the evaluator to go beyond the traditional emphasis on end-of-program performance and to make follow-up studies, based on ultimate objectives, a part of all program evaluations.

Provus also suggested an optional fifth stage that called for a cost-benefit analysis and a comparison of the results with similar cost analyses of comparable programs. In recent times, with funds for human services becoming scarcer, cost-benefit analyses have become a part of many program evaluations.

The Discrepancy Evaluation Model was designed to facilitate the development of programs in large public school systems and was later applied to statewide evaluations by a federal bureau. A complex approach that works best in larger systems with adequate staff resources, its central focus is on identifying discrepancies to help managers determine the extent to which program development is proceeding toward attainment of stated objectives. It attempts to assure effective program development by preventing the activity from proceeding to the next stage until all identified discrepancies have been removed. Whenever a discrepancy is found, Provus suggested a cooperative problem-solving process for program staff and evaluators. The process called for asking the following questions: (1) Why is there a discrepancy? (2) What corrective actions are possible? (3) Which corrective action is best? This process usually required that additional information be gathered and criteria developed to allow rational, justifiable decisions about corrective actions (or terminations). This particular problem-solving activity was a new addition to the traditional objectives-oriented evaluation approach.

Though the Discrepancy Evaluation Model was one of the earliest approaches to evaluation, elements of it can still be found in many evaluations. For example, in Fitzpatrick's interview with David Fetterman, a developer of empowerment evaluation, on his evaluation of the Stanford Teacher Education Program (STEP), Fetterman uses the discrepancy model to identify program areas (Fitzpatrick & Fetterman, 2000). The fact that the model continues to influence evaluation studies 30 years later is evidence of how these seminal

approaches continue to be useful to evaluators long after their original authors have ceased to espouse them.

A Schema for Generating and Analyzing Objectives: The Evaluation Cube

Building on a concept developed by Hammond (1973), The Evaluation Center at Western Michigan University developed a three-dimensional framework for analyzing the objects of community-based youth programs. This approach can easily be modified to incorporate relevant dimensions for any objectives-oriented program. The cube (Dodson, 1994, p. 61) is reproduced as Figure 6.1.

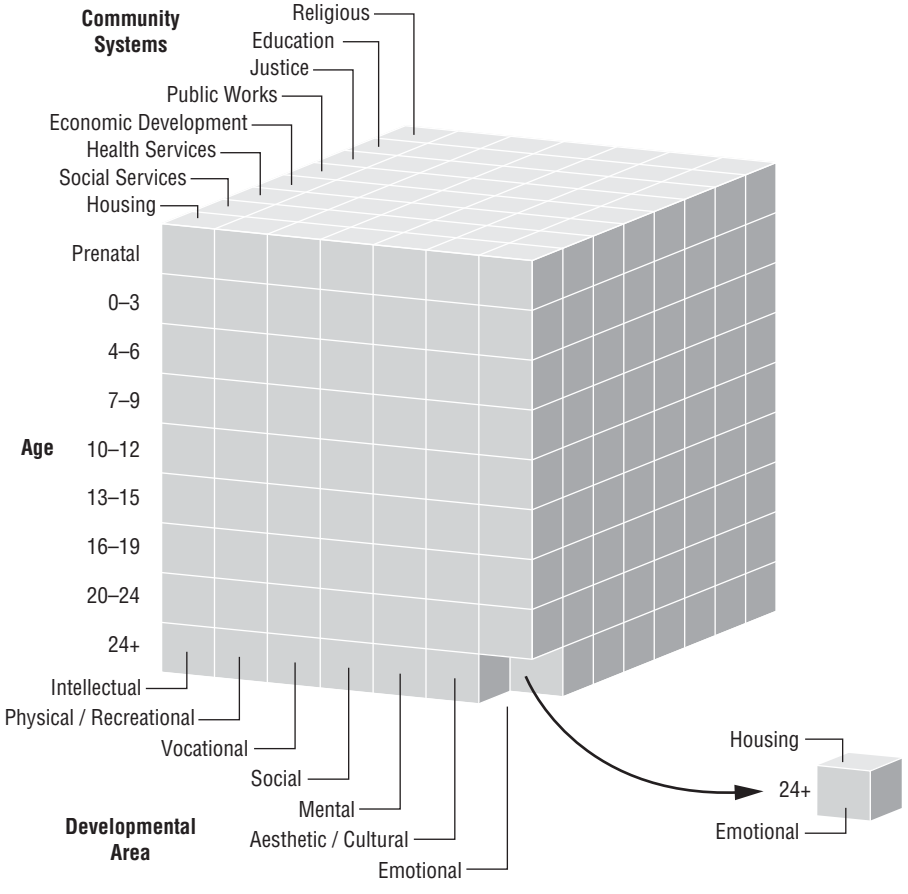


FIGURE 6.1 ***Objectives-Based Cube for Youth Program Analysis***

Source: From *Interim summative evaluation: Assessing the value of a long-term or ongoing program during its operations* (p. 58) by S. C. Dodson, 1994. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo. Reproduced by permission.

The three dimensions of the cube are as follows:

1. Needs of youth (the client): categories developed by Stufflebeam (1977) and expanded by Nowakowski et al. (1985) are
   - intellectual
   - physical recreation
   - vocational
   - social
   - moral
   - aesthetic/cultural
   - emotional
2. Age of youth (this dimension could be any relevant characteristic of the client): prenatal through young adult
3. Source of service to youth, such as
   - housing
   - social services
   - health services
   - economic/business
   - public works
   - justice
   - education
   - religious organizations

In any category along any of the three dimensions, those planning a community-based youth program may choose to establish relevant objectives. Few, if any, stakeholders in community-based programs will be interested in every cell of the cube, but the categories contained in each of the three dimensions will provide a good checklist for making certain that important areas or categories of objectives are not overlooked. Obviously, use of the cube is not limited to community-based programs but could extend to other types of programs as well.

Logic Models and Theory-Based Evaluation Approaches

Logic Models

One of the criticisms of objectives-oriented evaluation is that it tells us little about how the program achieves its objectives. This can be a particular problem when programs fail to achieve their objectives, because the evaluation can provide little advice on how to do so. Logic models have developed as an extension of objectives-oriented evaluation and are designed to fill in those steps between the program and its objectives. Typically, logic models require program planners or evaluators to identify program inputs, activities, outputs, and outcomes, with outcomes reflecting longer-term objectives or goals of the program and outputs

representing immediate program impacts. The model, typically presented in a diagram form, illustrates the logic of the program.

A typical logic model may include the following:

Inputs—annual budgets, staffing facilities, equipment, and materials needed to run the program

Activities—weekly sessions, curriculum, workshops, conferences, recruitment, clinical services, newsletters, staff training, all the key components of the program

Outputs—numbers of participants or clients served each week, number of class meetings, hours of direct service to each participant, number of newsletters and other immediate program products

Immediate, intermediate, long-term, and ultimate outcomes—the longitudinal goals for participant change (development)

Logic models are widely used in program planning and evaluation today. They have influenced evaluation by filling in the "black box" between the program and its objectives. Evaluators can use logic models to help program staff articulate and discuss their assumptions about how their program might achieve its goals and what elements are important to evaluate at any given time and generally to build internal evaluation capacity or the ability to think in an evaluative way. (See Taylor-Powell & Boyd [2008] for an example of the use of logic models in cooperative extension to build organizational capacity. Knowlton and Phillips [2009] also provide guidance for building logic models.) The United Way of America was one of the major organizations to bring logic models to evaluation through the logic-model-based approach it requires for the organizations it funds (United Way, 1996). Other foundations, such as the W.K. Kellogg Foundation and the Annie E. Casey Foundation, have also been instrumental in training organizations in the use of logic models to improve program planning and evaluation.

Theory-Based or Theory-Driven Evaluation

Carol Weiss first discussed basing evaluation on a program's theory in her 1972 classic book building on earlier writings by Suchman (1967) on the reasons that programs fail (Weiss, 1997; Worthen, 1996a). She has remained an effective and long-term advocate for theory-based evaluations (Weiss, 1995, 1997; Weiss & Mark, 2006). In the 1980s and 1990s, Huey Chen, Peter Rossi, and Leonard Bickman began writing about theory-based approaches to evaluation (Bickman, 1987, 1990; Chen & Rossi, 1980; 1983; Chen, 1990). Stewart Donaldson (2007) is one of the principal evaluators practicing and writing about the theory-driven evaluation approach today.[2] Edward Suchman (1967) had first made the point that programs can fail to

[2]Donaldson uses the term "theory-driven" but notes that the terms "theory-oriented," "theory-based," "theory-driven," and even "program logic" and "logic modeling" are all closely related or sometimes interchangeable. We use the terms "theory-driven" and "theory-based" interchangeably, but attempt to use the words used by the author we are discussing.

achieve their goals for two distinctly different reasons: (a) the program is not delivered as planned and, therefore, is not really tested (implementation failure); and (b) the program is delivered as planned and the results, then, clearly indicate that the program theory is incorrect (theory failure). He and Weiss recognized that, if an evaluation were examining whether a program achieved its goals and that program failed, it was important to know whether the failure was an implementation failure or a theory failure. With this information, the evaluator could then reach appropriate conclusions about the program and make useful recommendations for the decision maker. To distinguish between implementation failure and theory failure, the evaluator had to know two things in addition to simply measuring outcomes: (a) the essentials of the program theory. and (b) how the program was implemented. With this information, the evaluator could then determine whether the program implementation matched the theory. This was the beginning of program theory and the recognition of its importance to evaluation practice.

Chen's and Bickman's approaches to theory-based evaluation arose for these reasons, but also from their desire for evaluations to contribute more directly to social science research knowledge. Chen, for example, argued that evaluators of the time erred in focusing solely on methodology and failing to consider the theory or tenets of the program. For many of those writing about theory-based evaluation as it first emerged in the late 1980s and 1990s, *theory* meant connecting evaluation to social science research theories. Chen (1990), for example, encouraged evaluators to search the scientific research literature to identify social science theories that were relevant to the program and to use those theories in planning the evaluation. Evaluation results could then contribute to social science knowledge and theory as well as to program decisions (Bickman, 1987). Thus, theory-based evaluation arose from a science-based perspective and was often considered to be a strictly quantitative approach by others during the debates on qualitative and quantitative methods in the 1990s. However, today, theory-based evaluation is used by evaluators in many settings to gain a better understanding of the program. (See Rogers, 2000, 2001.) They can then use that understanding, the program theory, to better define the evaluation questions the study should address, to aid their choices of what concepts to measure and when to measure them, and to improve their interpretation of results and their feedback to stakeholders to enhance use.

But what is program theory? And what do evaluators using theory-based evaluation approaches do? Bickman defines program theory as "the construction of a plausible and sensible model of how a program is supposed to work" (Bickman, 1987, p. 5). More recently, Donaldson defines program theory as "the process through which program components are presumed to affect outcomes and the conditions under which these processes are believed to operate" (2007, p. 22). In both cases, and in other definitions, program theory explains the logic of the program. How does it differ from a logic model? In fact, they are quite similar. A logic model may depict the program theory if its articulation of program inputs, activities, outputs, and outcomes is sufficient to describe why the program is intended to achieve its outcomes. Logic models are sometimes used as tools to develop program theory. In other words, a program theory may look like a logic model. In our experience, because the emphasis in logic models is on the stages of

input, activities, outputs, and outcomes, the person developing the logic model focuses on listing all of the components in each of those categories and may, though certainly does not always, fail to describe the rationale or reasoning behind program success. In contrast, a program theory, while not containing prespecified categories such as inputs, activities, outputs, or outcomes, is intended to present the details of that reasoning. Bickman (1987) notes that program theory should clarify the relationship between the problem that the program assumes the student or client has and the program actions. Therefore, a program theory should begin with describing the client, or the assumptions the program has about the client, before the client begins the program. Logic models typically start a step after that, with program input. Another difference is simply that the organizations and people who write about logic models are different from those who write about program theory and theory-based evaluation.

Using Program Theory in Evaluation. The central element of theory-based evaluation approaches involves developing the theory for why the program should achieve its desired outcomes. It is this stage that distinguishes theory-based evaluations from other approaches. Chen (1990) differentiates between two models for developing program theory: (a) the stakeholder approach, in which the evaluator works with stakeholders, typically key program people, to discover their reasoning or underlying assumptions for program success; and (b) the social science approach, in which evaluators make use of their own knowledge both of the program and of the social science theory and research to develop a model.

Both Bickman and Chen note that developing program theory with stakeholders alone can often be problematic. Stakeholders may not know the program theory or, as Bickman writes, their theory may be "a vague notion or hunch" or "may be nothing more than a few simple assumptions about why the program should work" (1987, p. 6) because they are not well trained in social science theory or research. Or they may be purposefully obtuse or vague about program theory in order to obtain political support or funding or to avoid alienating an important group. Weiss, more recently, argues that improving the quality of program theory is one of the key challenges to evaluators. She notes that program managers or policymakers may or may not be able to develop a good-quality program theory. Program theories, she notes, must articulate the causal linkages between program actions and goals; if they do not, the theory is simply a model for implementation, a description of program processes. And, like Bickman, she expresses concern with relying too much on program stakeholders, planners, and practitioners to articulate that theory. She observes, "Many of these theories are elementary, simplistic, partial, or even outright wrong" (Weiss, 1997, p. 78), and emphasizes the need for evaluators to combine stakeholder input with social science research to build sound program theory.

The process of developing program theory should, therefore, rely on a combination of input from stakeholders, theories and research from relevant social science studies, and the evaluators' knowledge and expertise. Donaldson (2007), in his book on theory-based evaluation, decribes the steps for theory development

more precisely than early writers, based on his experience in using theory-based evaluation in several large projects. His steps reflect this balance:

1. *Engage relevant stakeholders.* The evaluator talks with as many representatives as possible from different constituencies to get their views on the program, its intended long-term outcomes, and the process the program uses to achieve those outcomes.

2. *Develop a first draft of program theory.* This step is undertaken by the evaluator or the evaluation team.

3. *Present the draft to stakeholders for further discussion, reaction, and input.*

4. *Conduct a plausibility check.* Now, the evaluators consult existing research and evaluations that are relevant to the program theory to assess the plausibility of each link. Does research suggest that the link could occur as planned? That the program action could lead to the intended outcome?

5. *Communicate these findings to key stakeholders and revise the program theory as needed.* Donaldson observes that the plausibility check may suggest that serious program changes are needed or that stakeholders have been overly optimistic about the outcomes that might be achieved. Evaluators present research findings to stakeholders and work with them to revise the program theory and/or the program itself so that the model accurately represents what will be done and what can be achieved.

6. *Probe arrows for model specificity.* As in step 4, the evaluators take the lead in examining the program theory "at a deeper level of detail." Donaldson indicates that at this stage the evaluators' focus is typically on critical links and discussion of details such as the length of time required for the outcome to occur and the nature of the process. The purpose, here, is for the evaluation team to ensure that they have an accurate, in-depth understanding of the program as it is intended to proceed. Such understandings can influence, for example, when evaluators collect data on outcomes and the type of data they collect.

7. *Finalize program impact theory.* The stakeholders have the final say in approving the model that will serve as the foundation for studying the program. Donaldson notes that he prefers relatively parsimonious models, as do some stakeholders, but others prefer models with more detail (Donaldson, 2007, pp. 33–39).

This theory-development process precedes any decisions about the evaluation. In fact, Donaldson indicates that stakeholders often want to move into thinking about the evaluation and the implications of the program theory during the process of discussing the evaluation. However, it is important for the program theory to be fully developed before moving into identifying the evaluation questions to be answered or the methods to be used to address such questions. The program theory, its key principles, should not be influenced by how the evaluation will be conducted—for example, by worrying about how certain linkages would be tested—but should instead reflect a true picture of what the program is intended to do and how it will do it.

The second phase of theory-driven evaluation is similar to that in most other evaluations. Working with stakeholders, the evaluators identify the key questions to be answered in the evaluation and the appropriate designs and methods for answering those questions. Although the theory-based evaluation approach has been considered a more quantitative approach because its proponents often come from quantitative areas, the approach does not prescribe or proscribe specific methods or designs. Often the emphasis is on testing the program model, that is, on questions of causality. The constructs identified in the program theory provide guidance as to what to measure and when to measure it (Lipsey, 1993). As Donaldson indicates, the selection of the evaluation questions to be addressed depend very much on the stage of the program (young versus mature enough to examine long-term outcomes) and what the stakeholders hope to learn. The development of program theory helps the evaluator to learn more about the program and its assumptions and, therefore, provides the evaluator with critical information that can be used throughout the evaluation. This includes determining what to study at different stages, identifying the constructs of interest, interpreting results, and making recommendations.

A typical model for theory-based evaluation would be to first study program implementation, focusing on whether key elements of the program theory are, in fact, delivered as planned. If so, the evaluator can then go on to study program outcomes knowing that this evaluation will be a test of program theory. If the program fails, it will mean that the theory does not work, at least with this client group in this context. But if the program is not implemented as planned, the evaluator may recommend changing the implementation to match the model, discarding the model as not feasible in this context, or trying some other model. In any case, the evaluation would not move on to measure outcomes because the program theory had not, in fact, been implemented. If outcomes were studied and success was achieved, the implementation study would demonstrate how the program that was delivered differed from the program theory. This modified program delivery might then become the standard model or program theory for the future.

Thus, the theory-based or theory-driven approach overcomes some of the failures of the objectives-oriented approach. It provides the evaluator with a way to look inside the black box and better understand what is happening between the time a student or client begins a program and when he or she concludes the program. With this information, theory-based evaluators argue, they can better test and determine the reasons for a program's success or failure.

How Program-Oriented Evaluation Approaches Have Been Used

The objectives-oriented approach has dominated the thinking and development of evaluation since the 1930s, both in the United States and elsewhere (Madaus & Stufflebeam, 1989). Its straightforward procedure of using objectives to determine a program's success or failure and to serve as a foundation for program

improvements, maintenance, or termination of program activities has proved an attractive prototype.

In education, the approach influenced the development of taxonomies of educational objectives (Bloom, Hastings, & Masia, 1971), the criterion-referenced testing movement of the 1960s and 1970s, and today's standards-based movement. As we noted in Chapter 2 in describing current trends in evaluation, the focus of evaluation today is on measuring outcomes; in schools, that takes the form of educational standards. No Child Left Behind (NCLB), the legislation passed by Congress in 2001, required all states that had not already done so to develop rigorous content standards for learning and tests to be given annually to measure accomplishment of those standards. This objectives-oriented means of evaluation now dominates K–12 education. Annual measurable objectives (AMOs) are used as a means of measuring progress toward the standards.

The objectives-oriented tradition has also influenced evaluation and management practices from the 1960s when Robert McNamara and the Rand Corporation brought Planning, Programming, and Budgeting Systems (PPBS) to the U.S. Defense Department, Management by Objectives (MBO), outcome monitoring (Affholter, 1994), and the Government Performance and Results Act (GPRA) (National Performance Review, 1993). Today, the Program Assessment Rating Tool (PART) has replaced GPRA (Office of Management and Budget, 2004), and performance monitoring has become a mainstay of many government management systems. Dahler-Larson (2006), in commenting on trends in evaluation today, sees performance monitoring as today's version of objectives-oriented or goal-oriented evaluation approaches. He observes that in-depth, goal-oriented evaluations have been replaced by monitoring of performance indicators. Performance monitoring systems are used by managers to monitor progress toward results. The systems may look at outputs, productivity, efficiency, service quality, customer satisfaction, or outcomes, but the focus is on the program and on the results (Positer, 2004). Logic models are sometimes used to identify the critical elements in the system that should be monitored, but to make it feasible to monitor these elements in an ongoing way requires compromises. So, data for performance monitoring systems tend to be solely quantitative and cost-effective to collect. Rarely are true long-term outcomes measured in an ongoing performance monitoring system if those outcomes are at all complex.

Although versions of objectives-oriented approaches continue to be popular with many government agencies and foundations for intensive evaluations, theory-based evaluation approaches are often the approach of choice by professional evaluators, particularly those with a more scientific bent. Many government funding agencies, particularly at the federal level in the United States, require programs to articulate their program theory or logic model. In addition, foundations such as the Aspen Institute, with their work on comprehensive community initiatives designed to have an impact at the community level, have pursued theory-based evaluations as a way to help them articulate the theory of complex programs and, then, to evaluate that theory as implemented (Weiss, 1995).

Strengths and Limitations of Program-Oriented Evaluation Approaches

Probably the greatest strength and appeal of the objectives-oriented approach lies in its simplicity. It is easily understood, is easy to follow and implement, and produces information that program directors generally agree is relevant to their mission. This approach has caused program directors to reflect about their intentions and to clarify formerly ambiguous generalities about intended outcomes. Discussions of appropriate objectives with the community being served have given objectives-oriented evaluation the appeal of face validity—the program is, after all, merely being held accountable for what its designers said it was going to accomplish, and that is obviously legitimate.

Useful as this approach to evaluation seems to its many adherents, it has many drawbacks. Principal ones include the single-minded focus on objectives and their measurement. The focus on objectives can cause evaluators to ignore other important outcomes of the program, both beneficial and detrimental, and if the evaluation draws final conclusions, the judgment of the program may be seriously incomplete. The evaluator may focus on the objectives like a horse wearing blinders. He or she may ignore the road to the right that drops off precipitously or the absolutely breathtaking view to the left in its efforts to look at (and measure) only the goal to be reached—the objectives that have been articulated. The objectives-oriented approach also neglects program description, the need to gain an understanding of the context in which the program operates and the effects of that context on program success or failure. Finally, evaluators using this approach may neglect their role in considering the value of the objectives themselves. Are these objectives, in fact, important ones for the program and its clientele? The ethical principles of evaluation, in Guiding Principle E, require the evaluator to consider "not only the immediate operations and outcomes of whatever is being evaluated but also its broad assumptions, implications, and potential side effects" (American Evaluation Association, 1995, p. 25). The objectives-oriented approach can appear seductively simple to novice evaluators who are only partially familiar with its philosophical and practical difficulties. Choices are involved in deciding which objectives to evaluate and how to interpret success or failure in each.

In today's standards-based environment, evaluators have little authority over state-required tests to measure standards. However, evaluations concerning standards should help stakeholders consider which standards are appropriate for their students and what levels, when reached, will be considered a success. If test items do not fully reflect the objectives of a particular school or district, results from alternative measures can be provided as support for achievement to parents and community leaders. Such evaluations can open up discussions about standards and goals in education for different communities.

Evaluations that make use of logic models or program theory to learn more about the program and to shed light on what to evaluate and the appropriate means for doing so obviously overcome some of the criticisms to the objectives-oriented

approach. Evaluators engage the stakeholders in dialogue so that they can learn more about the program, begin to develop a relationship with the stakeholders, and thereby gain a better understanding of what the evaluation might do. By the time a good theory-based evaluator moves to planning the evaluation, she should have a much better understanding of the values and concerns of the stakeholders regarding the program and the evaluation. The requirement for dialogue, as well as achieving an understanding and a clear articulation of the reasoning behind the program, is an obvious advantage of logic models and theory-based evaluations. These evaluators have focused on the program to identify and formulate the questions the evaluation will address, the timing of data collection, and the appropriate methods to be used. Some theory-based evaluators may feel obligated to evaluate the entire program theory, but that comprehensive focus is not the point of the process. The point is to understand the program from beginning to end and, then, to choose the appropriate links or components to evaluate given the stage of the program and the information needs of the stakeholders. Nevertheless, theory-based evaluators, like objectives-oriented evaluators, may struggle to see beyond their self-imposed blinders. Of course, theory-based evaluators are likely to focus on the theory and may ignore unintended program actions, links, outputs, or outcomes that merit attention. Further, their desire to test the theory as a whole may prompt them to neglect values or information needs of stakeholders. (See Fitzpatrick & Donaldson [2002] as Donaldson describes the pressures to evaluate the program theory for a training program for the unemployed that had been validated in Detroit, but now moved to a different population in California.)

Theory-based approaches are also criticized for oversimplifying the complexity of program delivery and context (Pawson, 2003). The reality of delivering programs is complex and, certainly, program theories simplify that complexity. But that is the purpose of theories or models—to reduce the messiness and complexity of actual program delivery to a parsimonious model that identifies the key assumptions or critical elements necessary for program success. In that way, the model helps the evaluator identify the most important elements or linkages to evaluate. Nevertheless, such reductionism does fail to convey the complexity of the real program and describing that complexity can also be an important role of evaluation. Oversimplification often leads citizens and policymakers alike to fail to understand how difficult, and costly, it is for programs or schools to achieve stated goals. Dahler-Larson calls for the need "to bridge the gap between conventions for relatively simple representations of causal models, on the one hand, and complex reality on the other" (2006, p. 152). He argues that theory-based evaluators should attempt to develop "different representations of program theories which will be fruitful in various ways depending on the purpose of the evaluation" (2006, p. 152).

We close our discussion of limitations of these types of evaluation approaches by briefly describing Scriven's goal-free evaluation. His concerns with the limitations of objectives-oriented approaches led him to develop his now widely known proposals for goal-free evaluation (1972), still discussed today to make evaluators aware of the bias that a focus on particular program elements can impose. Although intentionally the opposite of objectives-oriented approaches, it seems logical to discuss this proposal here.

Goal-Free Evaluation

The rationale for goal-free evaluation can be summarized as follows: First, goals should not be taken as given. Goals, he argues, are generally little more than rhetoric and seldom reveal the real objectives of the project or changes in intent. In addition, many important program outcomes are not included in the list of original program goals or objectives. Scriven (1972) believes that the most important function of goal-free evaluation is to reduce the bias that occurs from knowing program goals and, thus, to increase objectivity in judging the program as a whole. In objectives-oriented evaluation, an evaluator is told the goals of the program and is, therefore, immediately limited in her perceptions—the goals act like blinders, causing her to miss important outcomes not directly related to those goals.

For example, suppose an evaluator is told that the goals of a dropout rehabilitation program are to (1) bring school dropouts into a vocational training program, (2) train them in productive vocations, and (3) place them in stable jobs. She may spend all her time designing and applying measures to look at such things as how many dropouts have been recruited into the program and how many have been placed and have remained in paying jobs. These are worthwhile goals, and the program may be successful on all these counts. But what about the fact that the crime rate of others (non-dropouts) who are receiving employment training has tripled since the dropouts were brought into the vocational training program? Indeed, a hidden curriculum seems to have sprung up: stripping cars. This negative side effect is much more likely to be picked up by the goal-free evaluator than by the objectives-oriented evaluator working behind her built-in blinders.

The following are major characteristics of goal-free evaluation:

- The evaluator purposefully avoids becoming aware of the program goals.
- Predetermined goals are not permitted to narrow the focus of the evaluation study.
- Goal-free evaluation focuses on actual outcomes rather than intended program outcomes.
- The goal-free evaluator has minimal contact with the program manager and staff.
- Goal-free evaluation increases the likelihood that unanticipated side effects will be noted.

It might be helpful to point out that objectives-oriented and goal-free evaluations are not mutually exclusive. Indeed, they supplement one another. The internal staff evaluator, of necessity, conducts a goal-directed evaluation. She can hardly hope to avoid knowing the goals of the program, and it would be unwise to ignore them even if she could. Program managers obviously need to know how well the program is meeting its goals, and the internal evaluator uses goal-directed evaluation to provide administrators with that information. At the same time, it is important to know how others judge the program, not only on the basis of how well it does what it is supposed to do, but also on the basis of what it does in all

areas on all its outcomes, intended or not. This is a task for the external goal-free evaluator who knows nothing of the program goals. Thus, goal-directed evaluation and goal-free evaluation can work well together. And, while the major share of a program's evaluation resources should not go to goal-free evaluation, it is unfortunate when all resources go to goal-directed evaluation on a program when the stated goals do not even begin to include all of the important outcomes.

Major Concepts and Theories

1. The objectives-oriented evaluation approach was one of the first approaches to evaluation and is still commonly used today. Some of its present-day forms are standards-based testing and accountability in education and performance monitoring systems used in many government programs.

2. The objectives-oriented approach focuses on articulating the objectives of a program and collecting data to determine the extent to which they are achieved. Ralph Tyler and Malcolm Provus were early advocates of different facets of objectives-oriented evaluation.

3. Logic models are often used by program managers and evaluators today to link program inputs, processes, outputs, and outcomes and can serve as the foundation for making decisions about program or evaluation activities.

4. Theory-based or theory-driven evaluation approaches make use of social science theories and research relevant to the program and stakeholders' assumptions about why the program should work, to develop program theories. These program theories then serve as a foundation for selecting evaluation questions and making decisions about what to study and when to collect data on it. Theory-based evaluations are a frequently used approach in evaluation today.

5. The goal-free evaluation approach was proposed primarily to identify any unanticipated side effects of a program that a goal-directed or objectives-oriented evaluation might miss because of the focus on the intended program outcomes rather than on the actual program outcomes.

Discussion Questions

1. What are some of the primary reasons for using an objectives-oriented evaluation approach? Name a program or policy that you think would be useful to evaluate with this approach and discuss your rationale.

2. How does Provus's discrepancy model add to your thinking about what evaluation might do? Can you think of a program or policy that might be usefully evaluated with this approach?

3. What do you see as the advantages and disadvantages to the theory-based approach?

4. You only have a finite amount of money to spend on evaluation in your organization, but you're free to use it in any way you would like. Would you rather use it for an

on going performance monitoring system or for two or three in-depth evaluation studies each year that focus on particular programs or problems?

5. All program-oriented approaches share the strength of focusing the evaluator on some key characteristics of a program. They also share the weakness of perhaps causing the evaluator to focus unduly on those elements and, hence, to ignore other program effects, desirable or undesirable. Do you think the strength is worth the drawbacks? That is, would you rather focus on specific program characteristics (objectives or elements identified in a logic model or program theory) or would you rather try a form of goal-free evaluation, evaluating the program with no knowledge of its intended purposes? Justify your choice.

Application Exercises

1. Cheryl Brown is a program administrator for a state department of social services. She has been responsible for implementing a parenting program in which the goal is to reduce the incidence of child abuse and child neglect. To evaluate the program, she decides that she will depend on one performance measure—the number of cases of reported abuse and neglect. Use what you have just learned about Tyler's approach to evaluation, Provus's Discrepancy Evaluation Model, logic models, and goal-free evaluation to expand the evaluation design for this program. What are the risks of basing the evaluation on just this one measure? What are the advantages?

2. Jane Jackson is a leader of the English faculty of Greenlawn Middle School, which is under considerable pressure to reduce the achievement gap in writing that exists between Caucasian and minority students. It is May and she and the other English teachers are meeting to consider what data might help them examine their writing program in the next year and reduce the achievement gap in writing. Using what you have learned about different approaches in this chapter, advise Ms. Jackson on their evaluation. What elements of the approaches you have learned about might they use? How could she organize her evaluation? In particular, how might discussion of objectives, discrepancies, cubes, logic models, or program theory help them?

3. Many school districts and government agencies today are considering some form of performance-based pay for their employees. You are preparing to meet with a school district to discuss an evaluation of its performance-based pay program. You are considering using a theory-based approach to learn more about the key assumptions of this program. What do you think would be the advantages and disadvantages of using this approach to evaluate the program? With your classmates, develop a possible theory for such a program.

Case Studies

Three cases make use of theory-based evaluation and discuss some aspects of how they developed program theory: *Evaluation in Action*, Chapters 4 (Len Bickman), 9 (Stewart Donaldson), and 12 (Katrina Bledsoe).

In Chapter 4, Bickman discusses his measurement of program implementation before assessing outcomes to ensure that the model was being implemented. He also discusses his evaluation of the quality of that implementation,

something he believes we often neglect. The journal source is Fitzpatrick, J. L., & Bickman, L. (2002). Evaluation of the Ft. Bragg and Stark County systems of care for children and adolescents: A dialogue with Len Bickman. *American Journal of Evaluation*, *23*, 67–80.

In Chapter 9, Donaldson discusses how he uses program theory in several different cases to guide evaluations, though he encounters resistance in one. The journal source is Fitzpatrick, J.L., & Donaldson, S.I. (2002). Evaluation of the Work and Health Initiative: A dialogue with Stewart Donaldson. *American Journal of Evaluation*, 23, 347–365.

In Chapter 12, Bledsoe illustrates working with her colleague and the staff of a small program to develop program theory and how this process influences them and her. The journal source is Fitzpatrick, J.L., & Bledsoe, K. (2007). Evaluation of the Fun with Books Program: A dialogue with Katrina Bledsoe. *American Journal of Evaluation*, 28, 522–535.

Suggested Readings

Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications.* New York: Lawrence Erlbaum Associates.

Frechtling, J. A. (2007). *Logic modeling methods in program evaluation.* San Francisco: Jossey-Bass.

Tyler, R. W. (1991). General statement on program evaluation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century*. Ninetieth yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press.

United Way of America. (1996). *Measuring program outcomes*. Alexandria, VA: United Way of America.

Weiss, C. H., & Mark, M. M. (2006). The oral history of evaluation Part IV: The professional evolution of Carol Weiss. *American Journal of Evaluation, 27*(4), 475–483.

7

Decision-Oriented Evaluation Approaches

Orienting Questions

1. Why did decision-oriented evaluation approaches emerge?
2. What are the developmental stages of a program and how can decision-oriented evaluation help at each stage?
3. What is the personal factor? What are other key factors of utilization-focused evaluation?
4. What is performance monitoring? How is it like and unlike evaluation?
5. What are major strengths and limitations of the decision-oriented evaluation approaches as a group? As individual approaches?

Decision-oriented evaluation approaches were designed to address the problems that evaluations encountered in the 1970s—being ignored and having no impact. These approaches are meant to serve decision makers. Their rationale is that evaluative information is an essential part of good decision making and that the evaluator can be most effective by serving administrators, managers, policymakers, boards, program staff, and others who need good evaluative information. The three major decision-oriented approaches or methods we will review here are the CIPP model, which takes a systems approach to the stages of program development and the information needs that may occur at each stage; utilization-focused evaluation (UFE), which identifies primary users and works closely with them to identify information needs and conduct the study; and performance monitoring, which is not truly evaluation but provides information to managers to help in decision making and has been advocated by well-known evaluators. CIPP and

utilization-focused evaluation are rather different—the first is system and stage oriented, while the second is people oriented. But they share a firm goal of improving decision making in schools, nonprofits, and government. You will find elements of each that can be helpful in improving your own evaluations.

Developers of Decision-Oriented Evaluation Approaches and Their Contributions

Important contributions to the decision-oriented approach have been made by many evaluators. In education, Daniel Stufflebeam was a leader in developing an approach oriented to decisions. In the mid-1960s, Stufflebeam (1968) recognized the shortcomings of available evaluation approaches. Working to expand and systematize thinking about administrative studies and educational decision making, Stufflebeam (1968) made the decision(s) of program managers, rather than program objectives, the pivotal organizer for the evaluation. This made him one of the first evaluation theorists to focus on use. In the approaches proposed by him and other theorists (e.g., Alkin, 1969), the evaluator, working closely with an administrator, identifies the decisions the administrator must make, based on the stage of the program, and then collects sufficient information about the relative advantages and disadvantages of each decision alternative to allow a fair judgment based on specified criteria. The success of the evaluation rests on the quality of teamwork between evaluators and decision makers.

Michael Patton, with his utilization-focused approach, was another leader in focusing evaluations on decisions and use. In 1978, he published the first book on UFE. Patton argued that the first task of the evaluator was to identify a key user, often a manager with interest in the evaluation and with the authority and interest to make decisions with it.

The Decision-Oriented Approaches

The CIPP Evaluation Model

Stufflebeam (1971, 2004b, 2005) has been an influential proponent of a decision-oriented evaluation approach structured to help administrators make good decisions. He defines evaluation as "the process of delineating, obtaining, reporting and applying descriptive and judgmental information about some object's merit, worth, probity, and significance to guide decision making, support accountability, disseminate effective practices, and increase understanding of the involved phenomena" (Stufflebeam, 2005, p. 61). This definition expands on his original definition in 1973 when he first developed the CIPP model, but is essentially quite similar. Then, he defined evaluation more succinctly as "the process of delineating, obtaining, and providing useful information for judging decision alternatives" (Stufflebeam, 1973b, p. 129). The newer definition emphasizes the importance of

judging merit and worth, something that was central to evaluation in 1973. But his 2005 definition also emphasizes the currency of accountability, dissemination, and understanding in today's world of evaluation. However, the essentials of his CIPP model remain the same and, today, are used widely in the United States and around the world in educational evaluation. He developed this evaluation framework to serve managers and administrators facing four different kinds of decisions:

1. *Context evaluation,* to serve planning decisions: Determining what needs are to be addressed by a program and what programs already exist helps in defining objectives for the program. Context evaluation, as the name implies, concerns studying the context for a program that has not yet been planned: What are the needs and problems of students or clients? What assets or qualifications does the organization have to address these needs? What should be the goals and intended outcomes for a program?

2. *Input evaluation,* to serve structuring decisions: After defining needs and considering organizational assets and potential interventions, using input evaluation helps managers to select a particular strategy to implement and to resolve the problem and make decisions about how to implement it.

3. *Process evaluation,* to serve implementing decisions: Once the program has begun, the important decisions concern how to modify its implementation. Key evaluation questions are: Is the program being implemented as planned? What changes have been made? What barriers threaten its success? What revisions are needed? As these questions are answered, procedures can be monitored, adapted, and refined.

4. *Product evaluation,* to serve recycling decisions: What results were obtained? How well were needs reduced? What should be done with the program after it has run its course? Should it be revised? Expanded? Discontinued? These questions are important in judging program attainments.

The first letters of the four types of evaluation—context, input, process, and product—form the acronym CIPP, by which Stufflebeam's evaluation model is best known. Table 7.1 summarizes the main features of the four types of evaluation, as proposed by Stufflebeam (2005, p. 63).

As a logical structure for designing each type of evaluation, Stufflebeam (1973a) proposed that evaluators follow these general steps:

A. *Focusing the Evaluation*
   1. Identify the major level(s) of decision making to be served, for example, local, state, or national; classroom, school, or district.
   2. For each level of decision making, project the decision situations to be served and describe each one in terms of its locus, focus, criticality, timing, and composition of alternatives.
   3. Define criteria for each decision situation by specifying variables for measurement and standards for use in the judgment of alternatives.
   4. Define policies within which the evaluator must operate.

TABLE 7.1 ***The Relevance of Four Evaluation Types to Formative and Summative Evaluation Roles***

Evaluation Roles	*Context*	*Input*	*Process*	*Product*
Formative evaluation: Prospective application of CIPP information to assist decision making and quality assurance	Guidance for identifying needed interventions and choosing and ranking goals (based on assessing needs, problems, assets, and opportunities)	Guidance for choosing a program or other strategy (based on assessing alternative strategies and resource allocation plans), also for examining the work plan	Guidance for implementing the operational plan (based on monitoring and judging program activities)	Guidance for continuing, modifying, adopting, or terminating the effort (based on assessing outcomes and side effects)
Summative evaluation: Retrospective use of CIPP information to sum up the program's merit, worth, probity, and significance	Comparison of goals and priorities to assessed needs, problems, assets, and opportunities	Comparison of the program's strategy, design, and budget to those of critical competitors and the targeted needs of beneficiaries	Full description of the actual process and costs, plus comparison of the designed and actual processes and costs	Comparison of outcomes and side effects to targeted needs and, as feasible, to results of competitive programs; interpretation of results against the effort's assessed context, inputs, and processes

Source: From *Evaluation Theory, Models, and Applications* by D. L. Stufflebeam and A. J. Shinkfield. Copyright © 2007 by John Wiley & Sons. Reproduced with permission of John Wiley & Sons, Inc.

B. *Collection of Information*
   1. Specify the source of the information to be collected.
   2. Specify the instruments and methods for collecting the needed information.
   3. Specify the sampling procedure to be employed.
   4. Specify the conditions and schedule for information collection.

C. *Organization of Information*
   1. Provide a format for the information that is to be collected.
   2. Designate a means for performing the analysis.

D. *Analysis of Information*
   1. Select the analytical procedures to be employed.
   2. Designate a means for performing the analysis.

E. *Reporting of Information*
   1. Define the audiences for the evaluation reports.
   2. Specify the means for providing information to the audiences.
   3. Specify the format for evaluation reports and/or reporting sessions.
   4. Schedule the reporting of information.

F. *Administration of the Evaluation*
   1. Summarize the evaluation schedule.
   2. Define staff and resource requirements and plans for meeting these requirements.
   3. Specify the means for meeting the policy requirements for conducting the evaluation.
   4. Evaluate the potential of the evaluation design for providing information that is valid, reliable, credible, timely, and pervasive (that will reach all relevant stakeholders).
   5. Specify and schedule the means for periodic updating of the evaluation design.
   6. Provide a budget for the total evaluation program (p. 144).

Evolution of the CIPP Approach. The CIPP model has had the most staying power of any early evaluation model. Its principles have remained solid: the focus on serving decisions, judging merit and worth, the four stages of a program, reflecting the importance of context in considering evaluation questions, and an emphasis on standards and use. Its focus has traditionally been on program improvement. Stufflebeam, building on Egon Guba, writes, "evaluation's most important purpose is not to prove but to improve" (2004b, p. 262). He notes that his modification is not to exclude proving as a purpose, but to acknowledge that the primary purpose is improvement. With CIPP, Stufflebeam has always emphasized using multiple methods, both qualitative and quantitative—whatever methods are most appropriate for measuring the construct of interest.

Nevertheless, as Stufflebeam noted in 2004, "the CIPP model is a work in progress" (2004b, p. 245). The approach has been influenced by changes in evaluation practice and learning, as the model has been implemented in many different settings over the years. Although the original CIPP model focused very much on managers as the primary stakeholders, today's CIPP recommends involving many stakeholders, though the focus remains on decisions. The evaluator remains in firm control of the evaluation, but, Stufflebeam writes, "evaluators are expected to search out all relevant stakeholder groups and engage them in communication and consensus-building processes to help define evaluation questions, clarify evaluative criteria; contribute needed information; and reach firm, defensible conclusions" (2005, p. 62). Similarly, Stufflebeam is more forthright today in acknowledging that evaluation occurs in a political environment and that values play a key role. He writes, "Throughout my career, I have become increasingly sensitive to evaluation's political nature. Evaluators must regularly seek, win, and sustain power over their evaluations to assure their integrity, viability, and credibility" (2004b, pp. 261–262).

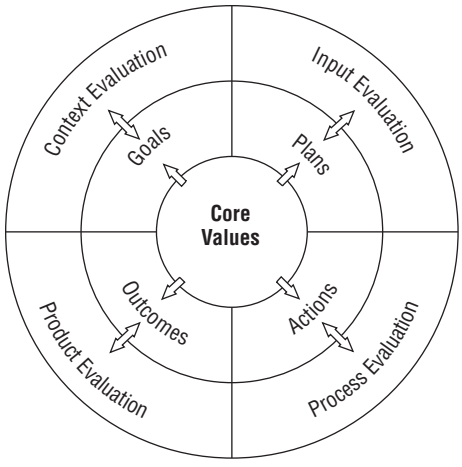


FIGURE 7.1 ***Key Components of the CIPP Evaluation Model and Associated Relationships with Programs***

Source: From *Evaluation Theory, Models, and Applications* by D. L. Stufflebeam and A. J. Shinkfield. Copyright © 2007 by John Wiley & Sons. Reproduced with permission of John Wiley & Sons, Inc.

Stufflebeam's wheel (See Figure 7.1) illustrates the impact of core values on each evaluation activity. The evaluation, he writes, should be grounded in these values which include "ideals held by society, group, or individual" and "provide the foundation for deriving and/or validating particular evaluative criteria" for judging the program or for making decisions and "provide the basis for selecting/constructing the evaluation instruments and procedures, accessing existing information," and other evaluation decisions (Stufflebeam, 2004b, p. 250).

Stufflebeam's work and his approach have added elements that differ from other approaches. His emphasis is practical, improving programs through improving decisions. He has written about and advocated many practical tools, including means for negotiating contracts, use of stakeholder panels for review and input, development of professional standards,[1] and metaevaluations—the evaluation of evaluations. He established the Evaluation Center at Western Michigan University whose web site includes many tools and checklists for evaluation approaches and tasks, including information on developing budgets, contracts, and negotiating agreements. See http://www.wmich.edu/evalctr/checklists/checklistmenu.htm

Significant Contributions of CIPP. Alkin and Christie, in their review of evaluation theories, use a tree with three main branches—use, methods, and valuing—to illustrate the many different evaluation theories. They place Stufflebeam at the root of the "use" branch and write that, "Stufflebeam's CIPP model is one of the most well-known of these [use] theories" (2004, p. 44). The CIPP approach has proved appealing to many evaluators and program managers, particularly those at

[1]Stufflebeam initiated the development of the Program Evaluation Standards discussed in Chapter 3 and served as the director of the Joint Committee on Standards for Educational Evaluation for many years. These Standards serve as a guide for judging the quality of evaluations and helping evaluators consider their priorities both in the United States and in many other countries.

home with the rational and orderly systems approach, to which it is clearly related. Perhaps its greatest strength is that it gives focus to the evaluation. Experienced evaluators know how tempting it is simply to cast a wide net, collecting an enormous amount of information, only later to discard much of it because it is not directly relevant to the key issues or questions the evaluation must address. Deciding precisely what information to collect is essential. Focusing on informational needs and pending decisions of managers limits the range of relevant data and brings the evaluation into sharp focus. This evaluation approach also stresses the importance of the utility of information. Connecting decision making and evaluation underscores the very purpose of evaluation. Also, focusing an evaluation on the decisions that managers must make prevents the evaluator from pursuing fruitless lines of inquiry that are not of interest to the decision makers.

CIPP was instrumental in showing evaluators and program managers that they need not wait until an activity or program has run its course before evaluating it. In fact, evaluation can begin when ideas for programs are first being discussed. Because of lost opportunities and heavy resource investment, evaluation is generally least effective at the end of a developing program. But today's emphasis on outcomes and impact has reduced evaluation's role at the planning stages. Nevertheless, particularly when purposes are formative, examining issues concerning context, input, and process can be helpful in identifying problems before they have grown and in suggesting solutions that will work better at achieving outcomes. For example, process studies may identify ways that teachers or other program deliverers are implementing a program, such as deviating from the intended activities because they are not working or are not feasible. Discovering these new methods, modifying the program model to conform to the new methods, and training others in them can help achieve program success.

Although the program stages used by CIPP indicate that the evaluation should focus on the stage of the program and that different questions arise at different stages, another advantage of the approach is that it encourages managers and evaluators to think of evaluation as cyclical, rather than project based. Like performance monitoring, evaluating programs at each stage can provide a "continual information stream to decision makers to ensure that programs continually improve their services" (Alkin & Christie, 2004, p. 44, analyzing the CIPP approach).

Nevertheless, as we will discuss further in our review of decision-making approaches, CIPP is not without its critics. Of principal concern is that, although the current model encourages participation from many stakeholders, the focus is typically on managers. Other stakeholders, who may not have explicit decision-making concerns, will necessarily receive less attention in defining the purposes of the evaluation, the means of data collection, and the interpretation of results.

The UCLA Evaluation Model

While he was director of the Center for the Study of Evaluation at UCLA, Alkin (1969) developed an evaluation framework that closely paralleled some aspects of the CIPP model. Alkin defined evaluation as "the process of ascertaining the

decision areas of concern, selecting appropriate information, and collecting and analyzing information in order to report summary data useful to decision-makers in selecting among alternatives" (p. 2). Alkin's model included the following five types of evaluation:

1. *Systems assessment*, to provide information about the state of the system (similar to context evaluation in the CIPP model)
2. *Program planning*, to assist in the selection of particular programs likely to be effective in meeting specific educational needs (similar to input evaluation)
3. *Program implementation*, to provide information about whether a program was introduced to the appropriate group in the manner intended
4. *Program improvement*, to provide information about how a program is functioning, whether interim objectives are being achieved, and whether unanticipated outcomes are appearing (similar to process evaluation)
5. *Program certification*, to provide information about the value of the program and its potential for use elsewhere (similar to product evaluation)

As Alkin (1991) has pointed out, his evaluation model made four assumptions about evaluation:

1. Evaluation is a process of gathering information.
2. The information collected in an evaluation will be used mainly to make decisions about alternative courses of action.
3. Evaluation information should be presented to the decision maker in a form that he can use effectively and that is designed to help rather than confuse or mislead him.
4. Different kinds of decisions require different kinds of evaluation procedures (p. 94).

Utilization-Focused Evaluation

Utilization-focused evaluation (UFE) is a well-known approach that is based on two assumptions: (a) the primary purpose of evaluation is to inform decisions; and (b) use is most likely to occur if the evaluator identifies one or more stakeholders who care about the evaluation and are in a position to use it. Patton calls the latter "the personal factor" and defines it as "the presence of an identifiable individual or group of people who personally care about the evaluation and the findings it generates" (2008a, p. 66). The personal factor is a central element to UFE. Patton first identified it as a factor critical to use in a study of use he conducted in the mid-1970s. In that study he interviewed evaluators and users of 20 federal health evaluations to learn the factors that contributed to the use of the evaluation. Patton and his colleagues had identified 11 potential factors from a review of the literature, such as methodological issues, political factors, and nature of the findings (positive, negative, surprising). They found that when asked about the single factor that most influenced use, two factors consistently emerged: political considerations and what Patton now calls the personal factor, the presence of an

individual or group who cared about the evaluation and its results. Patton's UFE approach has built on these findings, helping the evaluator identify these individuals and working closely with them to achieve use.

Patton defines UFE as "a process for making decisions and focusing an evaluation on intended use by intended users" (1994, p. 317). Similarly, in his most recent edition of *Utilization-Focused Evaluation*, he defines UFE as "evaluation done for and with specific intended primary users for specific, intended uses" (2008a, p. 37). His decision focus is further confirmed by his definition of evaluation:

> Program evaluation is undertaken to inform decisions, clarify options, identify improvements and provide information about programs and policies within contextual boundaries of time, place, values, and politics. (2008a, p. 40)

Although Patton sees UFE as a type of participatory approach, because of the focus on working with a key stakeholder or group of key stakeholders, he acknowledges that many place it among the decision-oriented approaches (Patton, 1994). We have placed UFE in this chapter because of its focus on an intended use, typically a decision. Patton does make use of intensive primary stakeholder involvement to achieve the intended use because, like Cousins and Earle (1992, 1995), Greene (1988), and others, he believes that involving stakeholders increases their sense of ownership in the evaluation, their knowledge of it, and, ultimately, their use of the results.

The first step in UFE concerns identifying the intended user or users—individuals who care about the study and its results. This step is, of course, central to achieving the personal factor. Given today's focus on networks and collaboration, Patton emphasizes that a careful stakeholder analysis, to identify the right stakeholders for the evaluation, is more important than ever. He suggests considering two important factors in identifying primary stakeholders: (a) interest in the study, and (b) power in the organization and/or power in the program or policy to be evaluated (Eden & Ackerman, 1998). Of course, the ideal stakeholder would be high on both, but a stakeholder with both interest and connections to others with power can be more useful than a powerful stakeholder with low or no interest. The latter may fail to attend important meetings, respond to messages, or participate in meaningful ways, thus harming the overall quality of the study and its credibility to others in the organization.

To help these primary users think about their needs for the evaluation, Patton indicates that he pushes users "to be more intentional and prescient about evaluation use during the design phase" (2008a, p. 146). He also suggests questions to ask these intended users to help them consider decisions and the feasibility of affecting them and the type of data or evidence that would be most likely to have an effect.

The remaining stages of UFE concern involving these stakeholders in the conduct of the study. This might include anything from identifying their questions of interest, which would then serve as the focus of the study, and considering how they would use the information obtained, to involving them in the

design and data collection stages, making sure they understand the methodology and that the choices made reflect their values and produce credible results that are useful to them. In the final stage, the primary stakeholders in UFE are involved in interpreting the results and making decisions about judgments, recommendations, and dissemination. The nature of the interaction between the evaluator and the primary intended users during these stages is very important in securing the personal factor. The evaluator is developing a personal relationship with the primary users to meet their needs and sustain their interest in the evaluation.

In many ways, these stages are similar to those in practical participatory evaluation approaches (PPE) such as those of Cousins and Earl (1992, 1995). The difference in the UFE model is the initial stage of stakeholder selection and the focus, at that stage, on intended, specific use. Although Cousins and Earl are selecting stakeholders to increase use, they consider the term *use* more broadly to include learning from participating in the process of evaluation (process use), conceptual use (knowledge gained that may be used in the future), and organizational learning. Similarly, Fetterman's empowerment evaluation considers self-determination and, more recently, ongoing organizational learning through establishing evaluation systems to be primary uses. In his focus on instrumental, or direct use, Patton's UFE is more like Stufflebeam's CIPP model, though CIPP focuses on the stage of the program and the decisions that would likely occur there, while Patton focuses on the decision makers and the dialogue with them to determine what decisions they think they will make. Patton's emphasis on the personal approach and relationships is also somewhat distinct from the CIPP approach.

Patton notes that the Achilles heel of UFE is staffing changes or turnover of the evaluation's primary intended users. As prevention strategies, he proposes having more than one primary user, ideally a task force of primary users, and building in enough time to brief replacements if the primary intended users are lost (Patton, 2008a).

Another criticism of the approach concerns Patton's emphasis on instrumental use by an individual or a small group and his view of how decisions are made. Carol Weiss and Michael Patton had some fiery debates in the 1980s on this issue. She thought Patton oversimplified decision making with his focus on a few primary users and unchanging contexts and decisions. Weiss today notes that the conventional wisdom, back when she started in evaluation in the 1960s and 1970s, was that there was one decision maker at the top and, "If you could get to that person with your data and conclusions, you could convince him—and it was always a 'him'—to switch from A to B. . . ." But, she argues, that's not the way organizations work:

> It is a much less tidy, much more back-and-forth, in-and-out, around-and-about kind of process, and all kinds of irrelevancies get tangled in the process as well. It's a complicated business and certainly not 'Find the decision maker and give him the word.' (Weiss & Mark, 2006, p. 480)

Studies on use find that it is complex, but, suggesting the value of Patton's approach, Alkin writes the following in a review of research on evaluation use:

> Perhaps the most influential of the evaluation factors identified by researchers is the evaluator. . . . Perhaps of greater importance than the evaluator's expertise and credibility are his or her personal characteristics, such as personality and style. Ability to develop rapport with users and to involve users in the evaluation is key if utilization is to occur. (2005, p. 455)

Another difference between Weiss' and Patton's views is the context of the evaluation. Weiss' work has been primarily with high-level government officials, such as members of Congress and the federal cabinet, who deal with many different issues and are too busy to become very involved, or very interested, in one particular evaluation. Patton's work is closer to the actual program being evaluated. Alkin (2005) notes that context is also an important factor in use, and one can readily see how these two different contexts would lead to different types of involvement of stakeholders and use of results.

Evaluability Assessment and Performance Monitoring

Joseph Wholey, like Michael Patton and Daniel Stufflebeam, has been prominent in the evaluation world for many years. Stufflebeam's work, however, lies mainly in education. Patton's is concerned with individual programs in schools and in social welfare settings. Wholey's influence and work has been with the federal government, starting with his work with the U.S. Department of Health, Education, and Welfare (HEW) in the 1970s. His focus is federal policy decision making. But like Patton and Stufflebeam, his goal is to have evaluation improve decisions. Therefore, he has developed several methods over the years to improve the utility of evaluation. We briefly review some of his major efforts here.

Evaluability assessment was developed by Wholey to prevent costly evaluations from being conducted when programs were not, in fact, ready for evaluation. Unlike Stufflebeam, who advocates evaluation during the context and input stages to help in program planning, Wholey's focus is typically on program outcomes (Wholey, 2004a, 2004b). In fact, most of his decision makers—federal policy makers—were not operating programs, and thus were not making formative decisions for program improvement; instead, they were making summative decisions regarding program funding, initiation, and continuation (M. Smith, 2005). Thus, Wholey's work at the federal level presents a stark contrast to the CIPP and UFE approaches, which are designed to work with policymakers and managers who are closer to the program being evaluated. During his early work with HEW, he and his colleagues were concerned that many evaluations were not being used. One reason for this, they wrote, was that the people implementing the programs had not had the opportunity to work things out, to clearly define what they were doing, to try them out, and to consider what information they needed from an evaluation. Therefore, he proposed evaluability assessment to help improve the likelihood that when

evaluations were conducted, the program would, in fact, be ready for evaluation. In order to be ready for evaluation, the evaluability assessment was designed to determine the following:

1. Whether the program purposes and goals were clearly defined.
2. Whether the program activities could feasibly achieve the intended goals.
3. Whether program managers were interested in program improvement.

The evaluator then worked with program managers, observed the program, read materials, conducted interviews, and carried out other activities to determine if these criteria were met.

Evaluability assessment declined after Wholey left HEW (see Rog, 1985; M. Smith, 2005), but we have chosen to introduce it here because it illustrates another decision-oriented method used in a quite different context. Smith (1989) later developed and successfully implemented evaluability assessment at the U.S. Department of Agriculture, but notes that, today, its methods have been adapted by theory-based evaluation and participatory approaches (Smith, 2005). Thus, evaluability assessment laid the groundwork for the importance of understanding program theory (steps 1 and 2 of the process) and understanding the needs of decision makers (step 3). We still see evaluability assessment as a valuable tool and describe how to conduct one in Chapter 11.

Wholey has continued active work in evaluation since the 1960s and, today, his emphasis is far more formative. In 2004, perhaps reflecting on the early failures of summative evaluations at HEW, he wrote: "Because policy decisions are influenced by so many inputs, policies are implemented through programs and programs tend to endure. I am particularly interested in the use of evaluation to improve program performance" (2004a, pp. 267–268). His first sentence reveals his recognition that summative evaluations at the federal level rarely are successful at influencing decisions about program continuation. Too many other factors are at work. Therefore, his focus today, like those of Stufflebeam and Patton, is on formative evaluations.

Wholey's recent evaluation-related work concerns performance monitoring systems as a means for improving decision making in organizations. Performance monitoring systems routinely collect data on program outputs or outcomes. Unlike most evaluation studies, performance monitoring is ongoing. It is not based on a particular program or project, but is a system that collects, maintains, and analyzes performance data for managers with the intent of using it to serve decisions and improve organizational performance. Wholey sees performance monitoring and evaluation as "mutually reinforcing" (2004a, p. 268). Performance monitoring systems should encourage managers to become accustomed to using data. They should, then, be more receptive to evaluation studies on particular programs or initiatives. Evaluations, in fact, might be initiated when managers become troubled by disappointing data from performance monitoring systems. They may ask: Why are we not achieving our goals?

But performance monitoring systems can be problematic. They may divert attention and resources to performance indicators that are not particularly meaningful

for managers or program staff.[2] As Wholey notes, "Performance monitoring is especially useful when the performance indicators are developed, tested, and refined, with the participation of program managers, staff, and other key stakeholders and are used in results-oriented management systems" (2004a, p. 268). We introduce Wholey's work and his advocacy of performance monitoring in this chapter because it, too, is intended as a decision-based approach to lead to better program management. Because performance monitoring is such a common form of data use in schools and organizations today, we wanted to provide the reader with an understanding of its origins and how it fits within the categories of evaluation approaches. Although evaluation has grown primarily as a periodic, project-based activity, many evaluators have talked about evaluation as an ongoing process to achieve learning in organizations (Preskill & Torres, 1998; Torres & Preskill, 2001; Owen & Lambert, 1998). Performance monitoring, properly conducted, may be one tool in that process. (See Poisiter, 2004; Wholey, 1999a, 1999b, 2001, 2003.)

How the Decision-Oriented Evaluation Approaches Have Been Used

Each of the contemporary approaches described here—the CIPP model, UFE, and performance monitoring—have been used widely in the United States and Canada and around the world. The CIPP model has been used extensively in school districts and state and federal government agencies. A guide for school program evaluation following the CIPP steps has been published by Sanders and Sullins (2005). UFE also continues to be a popular model; Patton published the fourth edition of his book in 2008. Finally, as we noted in the introduction to this book, performance monitoring has become mandatory in many government agencies and schools. Use of performance monitoring for improvement, however, is uneven because often the data that are collected are not relevant to particular programs. It is used primarily for accountability rather than program improvement. However, data-based decision making is thriving in many school districts and organizations.

Strengths and Limitations of Decision-Oriented Evaluation Approaches

The decision-oriented approaches are among the oldest approaches to evaluation, but they are still in frequent use. People are still writing about them and using them as guides to designing individual evaluations or evaluation systems. Their longevity speaks to their success. Using various means—articulating the stages of

[2]The data systems of school districts and state departments of education could be viewed as performance monitoring systems, but often the data are collected in response to federal or state mandates and are not seen as a primary aid for internal decision making.

a program and potential information needs at each stage, identifying and describing the personal factor, or considering ongoing information needs—Stufflebeam, Patton, and Wholey have developed models that succeed because they provide information that helps people, typically managers or policymakers, make decisions. That was their intent, and in that they have succeeded.

One criticism of the decision-oriented approaches is, ironically, their focus on decisions. Although Stufflebeam has broadened the CIPP approach to involve many stakeholders, critics have argued that these approaches tend to neglect stakeholders with less power (House & Howe, 1999). Social equity and equality are not values directly addressed by decision-oriented models. Instead, these advocates might argue, their approaches are working to improve programs that may serve these stakeholders. Nonetheless, the focus on managers and their information needs could restrict the information that evaluators seek, the types of data they collect, and the dissemination of the results. If great care is not taken, the evaluator can become the "hired gun" of the manager and program establishment. But, as Stufflebeam emphasizes, the CIPP model is judged by its adherence to the *Standards* of the Joint Committee, and these emphasize broad stakeholder input. In both CIPP and UFE, the evaluator is in charge of the evaluation and, although focusing on the decisions of primary stakeholders, will often use advisory groups or seek information from other stakeholders to supplement decisions about the evaluation.

A potential weakness of these approaches is the evaluator's occasional inability to respond to questions or issues that may be significant—even critical—but that clash with or at least do not match the concerns and questions of the decision maker who is the primary audience for the study. In addition, programs that lack decisive leadership are not likely to benefit from this approach to evaluation.

Finally, these evaluation approaches assume that the important decisions and the information to make them can be clearly identified in advance and that the decisions, the program, and its context will remain reasonably stable while the evaluation is being conducted. All of these assumptions about the orderliness and predictability of the decision-making process are suspect and frequently unwarranted. The evaluator should be prepared to frequently reassess and make adjustments for change. As Patton has noted in his new work in developmental evaluation, organizational environments can be dynamical (Patton, 2009)!

Major Concepts and Theories

1. The major impetus behind the decision-oriented approach to evaluation was the failure of evaluations to be used in the 1970s. These models were developed to focus directly on decisions, typically the decisions of managers and policymakers, to increase use.

2. Stufflebeam's CIPP model describes four program stages (context, input, process, and product) and the types of decisions that managers or policymakers may face at each of those stages. Using the stages and suggested decisions, the evaluator works with program managers or a steering committee of stakeholders to determine the decision maker's concerns, informational needs, and criteria for effectiveness when developing the evaluation.

3. Michael Patton's UFE approach attempts to produce evaluations that improve decisions by using the personal factor—individuals who have an interest in the evaluation and the power to make decisions or changes as a result of it. UFE identifies these intended users and works closely with them to plan and implement the evaluation. The evaluator builds a personal relationship with the intended users, which enables her to understand their decisions and to design the evaluation to best address their decisions and their values and ways of understanding.

4. Joseph Wholey's evaluability assessment was designed to influence decisions about whether to conduct an evaluation by determining whether an evaluation could be useful at that time. It influenced some of today's theory-based and participative approaches. He advocates performance monitoring systems to provide ongoing output and outcome information to managers and to facilitate evaluation and data-based decisions.

Discussion Questions

1. Which decision-oriented approach do you find most appealing? Which would be most useful in your organization to improve decisions? Are there elements of the other approaches that you think would also be useful? Could you combine them?

2. Decision-oriented approaches focus on decisions that are intended to improve the use of the evaluation results and the impact of that use on the program. But the tradeoffs are that other groups may be ignored, evaluation concerns that don't lead directly to decisions might be neglected, and the decision makers, decisions, and even the context for the program may change. Is the focus on decisions worth the tradeoffs?

3. Almost every organization does some type of performance monitoring. Do you think it is useful, or should the resources used for collecting and maintaining that data go into individual evaluation studies of selected programs or policies?

4. The developers of the approaches described here emphasize that their primary intention is formative evaluation, or program improvement, but that they can also be used for summative decisions. Why, or in what circumstances, are these methods better for formative decisions? In what circumstances might the method be used for summative decisions?

Application Exercises

1. Using what you have just read about the decision-oriented approaches to evaluation, identify one or two decisions to be made about a program at your workplace or one with which you are familiar. Who are the decision makers? What information do you think they need to make the decisions? What information do they think they need to make the decisions? What will influence their decisions? Do you think an evaluation could help them in making these decisions? What kind of evaluation, using the CIPP model, would be most appropriate for each decision? (See Table 7.1 for types.)

2. Describe how decisions about programs are typically made in your organization. Would a decision-oriented approach work in your organization? Why or why not?

3. A public school district successfully demonstrated its need for federal support for a program to reduce the achievement gap between white and minority students in their elementary schools. They received a $1,000,000 grant to be spent over a period of three years from July 1, 2011, to June 30, 2014. On March 15, 2011, the superintendent convened a meeting of the assistant superintendent of elementary instruction and thirty principals of elementary schools eligible to participate in the proposed program. It was decided that a thorough evaluation of the current reading and mathematics programs in these schools should be completed by September 30, 2011, to identify needs. Alternative strategies for solving needs would then be investigated and a program chosen to reduce the achievement gap. They also decided to establish an evaluation team that would be responsible for the following:

- **a.** Conducting the evaluation of the reading and mathematics programs of the eligible schools
- **b.** Evaluating alternative programs to meet the needs of the thirty schools
- **c.** Continually monitoring the program, which would be implemented starting in 2012
- **d.** Collecting information to be reported annually (on June 30 for each year of the grant) to the U.S. Department of Education

Using what you have just learned about decision-oriented evaluation approaches, tell the evaluation team members how they should proceed (assuming that it is now March 2011). Be as detailed in your planning as you can be.

4. Identify an evaluation that was conducted in your organization or university and answer the following questions: Who initiated the evaluation (e.g., management, mandated by the federal government)? Did the evaluation use a particular approach? Who, other than the evaluation staff, was involved in the evaluation and how were they involved? What role did managers play? What was the purpose of the evaluation? What questions were answered by the evaluation? What types of data were collected? How did the primary stakeholders, and other stakeholders, receive the results? How did they react to the results? What was management's level of receptivity to the evaluation? Were the evaluation findings used? How? Next, consider whether management would have been more or less receptive to the evaluation and whether they would have used the evaluation findings differently if a decision-oriented approach had been employed. Discuss your answers.

5. Identify a central problem facing your organization or school, such as educating students, serving clients effectively, or recruiting and training new employees.

- **a.** Considering the CIPP model, at what stage would you consider the problem? Does considering the stage help you identify some potential information needs? If so, what questions should the evaluation address?
- **b.** Now apply the UFE approach to the same problem. Who would be the primary intended users? What decisions should they be considering? What information, that they do not currently have, do they need to make the decision? How could an evaluator work with them?
- **c.** Does your performance monitoring system or any routine data that you collect supply any useful information on the problem? If not, and if this is an ongoing problem, should you add something to the system to routinely collect data on the issue?

Case Studies

We recommend two interviews that concern decision making for this chapter: In *Evaluation in Action*, Chapters 2 (James Riccio) and 5 (David Fetterman).

In Chapter 2, Riccio's primary audience is the California state legislature, which makes decisions on welfare programs and has contracted for this evaluation to help them decide whether to continue the program. The audience—the legislature—is distant, but he discusses their influence and how he works with them. Their needs shape the nature of his evaluation. The journal source is Fitzpatrick, J. L., & Riccio, J. A. (1997). A dialogue about an award-winning evaluation of GAIN: A welfare-to-work program. *Evaluation Practice*, *18*, 241–252.

In Chapter 5, Fetterman, the developer of empowerment evaluation, chooses to use a decision-making approach rather than empowerment because of the information needs of his client, the new dean of education at Stanford University, and the client's history of decisions. His discussion of that choice makes for useful reading. The journal source for that chapter is Fitzpatrick, J. L., & Fetterman, D. (2000). The evaluation of the Stanford Teacher Education Program (STEP): A dialogue with David Fetterman. *American Journal of Evaluation*, *20*, 240–259.

Suggested Readings

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Sanders, J. R. & Sullins, C. (2005). *Evaluating school programs: An educator's guide* (3rd ed.). Thousand Oaks, CA: Corwin.

Stufflebeam, D. L. (2000). The CIPP model for evaluation. In D. L. Stufflebeam, G. F. Madaus, & T. Kelleghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed.), pp. 274–317. Boston: Kluwer.

Stufflebeam, D. L. (2004). The 21st-century CIPP model: Origins, development, and use. In M. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences.* Thousand Oaks, CA: Sage.

Wholey, J. S. (2001). Managing for results: Roles for evaluators in a new management era. *American Journal of Evaluation*, *22*(4), 343–347.

Wholey, J. S. (2003). Improving performance and accountability: Responding to emerging management challenges. In S. J. Donaldson & M. Scriven (Eds.), *Evaluating social programs and problems* (pp. 43–61). Mahwah, NJ: Lawrence Erlbaum.

8

Participant-Oriented Evaluation Approaches

Orienting Questions

1. Who were the historic leaders in establishing participant-oriented approaches and what did they contribute? What prompted their work?
2. How might we differentiate among the many different contemporary participative approaches?
3. How do the practical participative approaches differ from the transformative ones?
4. Contrast the Practical Participatory Evaluation, Empowerment Evaluation, Development Evaluation, and Deliberative Democratic Evaluation approaches. How do their purposes and methods differ? Under what conditions would you use each?
5. How are participatory approaches used in practice?
6. What are the major strengths and limitations of participant-oriented evaluation approaches?

Participant-oriented approaches to evaluation currently include many different models, but their commonality is that they all use stakeholders—people with an interest or "stake" in the program—to assist in conducting the evaluation. They may use them to achieve different goals: stakeholders' greater understanding and ownership of the evaluation leading to greater use of the results or to empowering stakeholders, building evaluation capacity in the organization, and increasing organizational learning and data-based decision making.

The approaches use stakeholders in quite different ways. Some approaches use them primarily at the beginning and ending stages of the evaluation to help

define the evaluation questions and, later, to interpret the findings and make recommendations. Others use stakeholders intensively throughout the process, sometimes having the stakeholders act as the major decision makers while the evaluator serves in the role of a technical consultant, as needed. In this chapter, we will describe how the earlier participant-oriented approaches emerged and the characteristics that continue to influence evaluation approaches today. Then, we will describe some of today's contemporary participatory approaches including their purposes, principles, and methods. Finally, we will discuss their strengths and weaknesses and how they might be used.

Evolution of Participatory Approaches

In the early years of evaluation in the United States, after it was first mandated by Congress and governmental agencies, most evaluation practitioners relied on traditional social science research methods to determine whether the goals and objectives of programs had been achieved and then provided that information to government policymakers. As early as 1967, though, several evaluation theorists began to react to what they considered to be the dominance of mechanistic and insensitive approaches to evaluation in the field of education. These theorists expressed concerns that evaluators were largely preoccupied with stating and classifying objectives, designing elaborate evaluation systems, developing technically defensible objective instrumentation, and preparing long technical reports, with the result that evaluators were distracted from what was really happening in the programs they were evaluating. Critics of traditional evaluation approaches noted that many large-scale evaluations were conducted without the evaluators ever once setting foot on the participating program site(s). What began as a trickle of isolated comments grew to a deluge that flooded evaluation literature in education and the social sciences. More and more practitioners began publicly to question whether many evaluators really understood the phenomena that their numbers, figures, charts, and tables were intended to portray. An increasing segment of the education and human services communities argued that the human element, reflected in the complexities of the everyday reality of the delivery of programs and the different perspectives of those engaged in providing services, was missing from most evaluations.

Consequently, a new orientation to evaluation arose, one that stressed first hand experience with program activities and settings and involvement of program participants, staff, and managers in evaluation. This general approach, which grew quickly after the early 1970s, was aimed at observing and identifying all (or as many as possible) of the concerns, issues, and consequences the different stakeholders had about the program being evaluated.

Due in large part to a reaction to perceived deficits in other evaluation approaches, this orientation now encompasses a wide variety of more specific

approaches that might be generally tied together by their acceptance of the constructivist paradigm, recognizing that there are many perspectives to knowledge and truth and, therefore, to a program and its evaluation (see Chapter 4.) Many of those who contributed to the development and use of participant-oriented approaches to program evaluation prefer naturalistic inquiry methods, as described later in this chapter. Moreover, most advocates of this approach see involving participants in the program, managers, staff, and other key stakeholders as a key principle of good evaluation—hence the descriptor "participant-oriented" as a label for this approach.

Developers of Participant-Oriented Evaluation Approaches and Their Contributions

Robert Stake and His Responsive Approach

Robert Stake (1967) was the first evaluation theorist to provide significant impetus to this orientation in the field of education. His paper, "The Countenance of Educational Evaluation," with its focus on portraying and processing the judgments of participants, dramatically altered the thinking of evaluators in the next decade. Along with his later development of responsive evaluation (Stake, 1973, 1975a, 1975b), he provided conceptions and principles that guided the evolution of this approach. Stake's early writings evidenced his growing concern over the dominance of program evaluation by what he saw as parochial, objectivist, and mechanistic conceptions and methods. Guba's (1969) discussion of the "failure of educational evaluation" then provided further impetus to the search for an alternative to the rationalistic approach to evaluation.

The Countenance of Evaluation. Stake's first departure from traditional evaluation was his development of the Countenance Framework (1967). In it, he asserted that the two basic acts of evaluation are description and judgment, the two *countenances* of evaluation. Evaluations should provide a full description of the program and entity being evaluated and then make a judgment of its merit or worth. To aid the evaluator in organizing data collection and interpretation, Stake created the evaluation framework shown in Figure 8.1.

Using the framework shown in Figure 8.1, the evaluator first determines the rationale for the program, which includes the need the program is intended to serve and relevant features of the program's development. (See Rationale box at far left.) The descriptive part of evaluation then focuses on first determining program intents (column 1) in regard to program antecedents (inputs, resources, and existing conditions), transactions (program activities and processes), and program outcomes. Through observations at each of these levels, the evaluator begins to thoroughly describe the program and to compare the program intents with the actual observations of the program in action. In the judgment stage, the evaluator explicitly identifies or develops standards (criteria, expectations, performance of

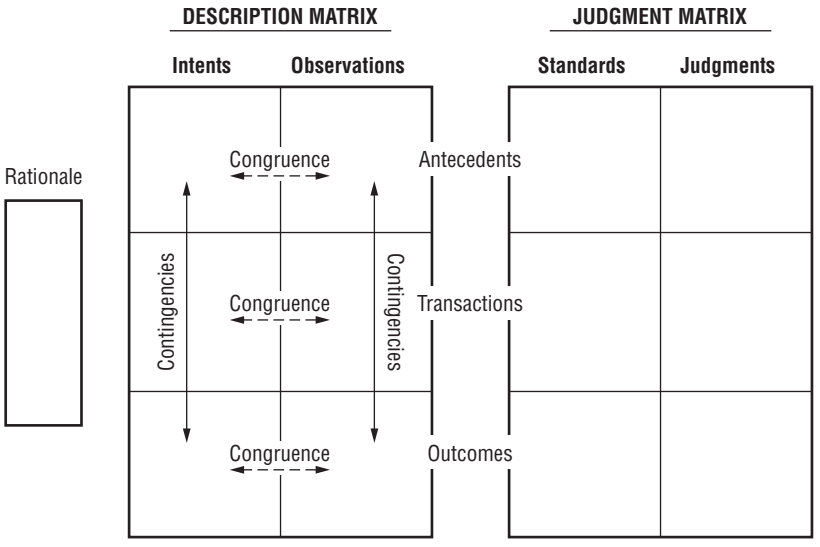


FIGURE 8.1 ***Stake's Layout of Statements and Data to Be Collected by the Evaluator of an Educational Program***

Source: From "The Countenance of Educational Evaluation" by R. E. Stake, 1967, *Teachers College Record, 68,* p. 529. Reprinted with permission.

comparable programs) for judging program antecedents, transactions, and outcomes and, finally, records judgments made about the antecedent conditions, transactions, and outcomes. The evaluator analyzes information in the description matrix by looking at the congruence between intents and observations, and by looking at the dependencies (contingencies) of outcomes on transactions and antecedents, and of transactions on antecedents. Judgments are made by applying standards to the descriptive data.

Thus, the countenance structure gave evaluators a conceptual framework for thinking through the data needs of a complete evaluation. In reviewing his countenance paper years later, Stake (1991) noted that it underemphasized the process of describing the evaluation, a shortcoming that he addressed later in his responsive evaluation approach. In fact, it was his descriptive emphasis that was new to evaluation approaches of the times. Stake wanted evaluators to become familiar with the particulars of the programs they were studying and to gain a thorough understanding of them before examining outcomes. The evaluator's understanding of the antecedents and the transactions would help him or her better interpret the successes or failures in achieving desired outcomes.

Responsive Evaluation. Stake's responsive evaluation, introduced in 1973, was more radical. In it, he truly tackled his concerns with the directions evaluation was taking at the time. Greene and Abma noted this in their foreword to a 2001 issue

of *New Directions for Evaluation* that focused on responsive evaluation, its influences, and its current applications and adaptations:

> Stake offered a new vision and rationale for educational and social program evaluation to the then-fledgling evaluation communities. In this vision, evaluation was reframed—from the application of sophisticated analytic techniques that address distant policymakers' questions of program benefits and effectiveness "on the average," to an engagement with on-site practitioners about the quality and meanings of their practice. These innovative ideas helped accelerate a transformation of the evaluation enterprise into its current pluralistic character, within which remain multiple and varied legacies of key responsive evaluation principles (2001, p. 1).

Although the seeds of this explication lie in his earlier work, Stake's subsequent conceptions of responsive evaluation (1973, 1975b, 1978, 1980) are implicitly less formal and explicitly more pluralistic and process focused than his earlier countenance model. The responsive evaluation approach departed from past evaluation approaches foremost in its flexibility and in its responsiveness to the particulars of the evaluation setting and its naturalness. Stake noted that he was not proposing a new approach to evaluation, for "responsive evaluation is what people do naturally in evaluating things. They observe and react" (1973, p. 1). Rather, Stake saw the responsive approach as a means to improve upon and focus this natural behavior of the evaluator. Stake stressed the importance of being responsive to realities in the program and to the reactions, concerns, and issues of participants, rather than being preordinate[1] with evaluation plans, relying on preconceptions, and setting formal plans and objectives of the program before gaining a full understanding of the program.

Stake defined responsive evaluation as follows:

> An educational evaluation is responsive evaluation if it orients more directly to program activities than to program intents; responds to audience requirements for information; and if the different value perspectives present are referred to in reporting the success and failure of the program. (1975a, p. 14)

Responsive evaluation differed from existing evaluation approaches in many ways that foreshadowed today's participatory approaches. These included:

(a) Flexible, changing methods and approaches; adapting to new knowledge as the evaluation proceeds; using an iterative, open-ended model.
(b) Recognition of multiple realities and the value of pluralism. Programs are seen by others in many different ways and the evaluator is responsible for portraying those many different pictures.
(c) Local knowledge, local theories, and the particulars of an individual program, its nuances and sensitivities, are more important to convey than testing any big theory or generalizing to other settings.

[1] "Preordinate" evaluation refers to evaluation studies that rely on prespecification, when inquiry tends to follow a prescribed plan and does not go beyond or vary from the predetermined issues and predefined problems.

(d) Case study and qualitative methods are important, essential methods to understand the particulars of a case and to correspond to the natural ways in which people come to understand something.
(e) Evaluations should strive to be holistic, to convey the full complexity of a program, not to reduce or simplify.
(f) Evaluation reports should follow this natural approach, presenting a rich set of information in full narratives emphasizing description and understanding.
(g) The evaluator may make a judgment, but his or her individual judgment may differ from those of others presented with the information; thus, the evaluator's role is also that of a learner and a teacher—a facilitator—to help others reach their own judgments.

The responsiveness and flexibility of the model are reflected in the *clock* Stake (1975b) developed to reflect the prominent, recurring events in a responsive evaluation. (See Figure 8.2.) Although the evaluator would typically begin the evaluation at twelve o'clock and proceed clockwise, Stake emphasized that any event can follow any other event, and at any point the evaluator may need to move counterclockwise or cross-clockwise, if events and increased understanding

FIGURE 8.2 ***Prominent Events in a Responsive Evaluation***

Source: From *Program Evaluation, Particularly Responsive Evaluation* (Occasional Paper No. 5, p. 19) by R. E. Stake, 1975b, Kalamazoo, MI: Western Michigan University Evaluation Center. Used with permission.

warrant such changes. Further, many events may occur simultaneously, and others will occur several times during an evaluation.

One revealing comparison of responsive and preordinate evaluation approaches was provided by Stake's (1975b) analysis of what percentage of their time evaluators of each persuasion would spend on several different evaluation tasks (p. 20):

	Preordinate (%)	***Responsive (%)***
Identifying issues, goals	10	10
Preparing instruments	30	15
Observing the program	5	30
Administering tests, etc.	10	—
Gathering judgments	—	15
Learning client needs, etc.	—	5
Processing formal data	25	5
Preparing informal reports	—	10
Preparing formal reports	20	10

A critical difference between traditional, social science-based evaluation and responsive evaluation that emerges in this comparison is the greater amount of time spent by the preordinate evaluator in preparing or developing instruments and in analyzing data. In fact, Stake surmises that these are the two dominant activities of the preordinate evaluator. In contrast, the dominant activity of the responsive evaluator should be observing the program, learning much more about what is actually going on in the program. Similarly, the responsive evaluator will spend more time than the preordinate evaluator gathering the judgments of others about the program, learning about clients' needs, and preparing informal reports.

Finally, Stake captures the responsive evaluator's role in this description:

> To do a responsive evaluation, the evaluator of course does many things. He makes a plan of observations and negotiations. He arranges for various persons to observe the program. With their help he prepares brief narratives, portrayals, product displays, graphs, etc. He finds out what is of value to his audience. He gathers expressions of worth from various individuals whose points of view differ. Of course, he checks the quality of his records. He gets program personnel to react to the accuracy of his portrayals. He gets authority figures to react to the importance of various findings. He gets audience members to react to the relevance of his findings. He does much of this informally, iterating, and keeping a record of action and reaction. He chooses media accessible to his audiences to increase the likelihood and fidelity of communication. He might prepare a final written report; he might not—depending on what he and his clients have agreed on. (1975b, p. 11)

Contributions and Commentary. Thus, Stake's responsive model introduced a new approach to evaluation, one that was more flexible and adaptive to the circumstances of the program and the needs of stakeholders. The emphasis was both on understanding through portraying in-depth descriptions and on the local

or particular issues of an individual program. Will Shadish, in discussing his review of Stake's work for his well-known book on the foundations of program evaluation (Shadish, Cook, & Leviton, 1991), notes that while studying Stake's work, he began to realize that Stake's advocacy of case study methods was not derived from his philosophical or paradigmatic preferences (Shadish & Miller, 2003). His reasons, Shadish argues, "were more political, more in terms of giving people control at the local level, empowering them" (2003, p. 268). Stake himself confirms Shadish's observations in an interview with Tineke Abma, stating, "I am a populist, a localist. I am fearful of central authority and of connoisseur control. . . . I am a situationalist, thinking that goodness in government, goodness in living, goodness in schooling is strongly dependent on the situation" (Abma & Stake, 2001, p. 10–11). Participative approaches today continue to have a local emphasis, though often a more overt political one concerned with empowering underrepresented stakeholders. We will discuss these approaches later in the chapter. Suffice it to say here that Stake's responsive model and advocacy for case studies was new to evaluation at the time and continues to influence the development of evaluation today.

His approach and his work were not without critics, although Shadish, Cook, and Leviton (1991) note that he is "virtually unique" among the founding evaluators in his localist, anti-federal, and anti-generalization approach. However, they question Stake's assumption that local stakeholders who have developed the interventions and programs that are being evaluated and who, therefore, have much stake in them, will use the information from the evaluation to change the program in dramatic or challenging ways that they may have resisted or not considered previously.

Another criticism concerned the evaluator's dominant role. The responsive approach differs from some of the participatory approaches described later in this chapter in that it does not advocate specifically for the inclusion of stakeholder groups who are underrepresented in the program decision-making process. House (2001), whose deliberative democratic approach will be reviewed in this chapter, notes that responsive evaluation does not address social justice or the inequalities among stakeholder groups and, in fact, is not participatory as we view participatory evaluation today. Stakeholders were not involved in conducting the evaluation. Instead, the evaluator remained in firm control of the evaluation. The responsive evaluator seeks the views of other stakeholders and attempts to portray the complexity and multiple views of the program, so he is, in that sense, participatory. That approach differed from the common outcome-oriented approaches of the day. But the approach was not participatory in the sense of involving stakeholders in decisions and actions of the evaluation as many current participatory models do. In clarifying this aspect of the approach, Stake more recently stated:

> With me, the locus of control needs to stay outside [with the evaluator], no matter how much I rely on them [stakeholders]. I do not advocate collaborative definition of research questions. To be responsive does not automatically yield design authority to stakeholders. It means coming to know the circumstances and problems and values well, then using professional talent and discipline to carry out the inquiry. For me, the inquiry belongs to the evaluator. (Abma & Stake, 2001, p. 9)

Thus, Stake's countenance model and responsive approach changed evaluation dramatically and opened the door to today's participant-oriented models. As the reader will see, these models differ from the responsive approach in important ways, both in how they involve stakeholders in the evaluation and in their goals for social justice or systems change. So why is Stake the historical antecedent to these models? Because he turned the existing models on their heads and introduced quite different methods. Collecting the views of different stakeholders and giving legitimacy to those views was new. Advocating for a local emphasis, to learn the particulars, the antecedents, and the processes, and to use qualitative methods to do so, was a huge change that gradually led evaluators to consider other ways to include stakeholders and purposes of evaluation.

Egon Guba and Yvonna Lincoln: Naturalistic and Fourth-Generation Evaluation

In the 1980s, Egon Guba and Yvonna Lincoln published two books that had a great influence on evaluation: *Naturalistic Inquiry* (1985) and *Fourth-Generation Evaluation* (1989). Like Stake, their original impetus was to move evaluation away from an emphasis on traditional, quantitative methods of data collection to consider qualitative, naturalistic methods, including interviews, observations, and case studies. However, their emphasis in developing naturalistic evaluation was much more philosophical and epistemologically based than Stake's. Guba, later joined by Lincoln, popularized the constructivist paradigm for evaluation. They focused on multiple realities and the need for the evaluator to construct those realities by collecting the input and views of different stakeholders.

Guba and Lincoln greatly advanced evaluators' philosophical thinking about the constructivist paradigm and its applications to evaluation. They developed new criteria for judging naturalistic or constructivist evaluations, some of which paralleled, but provided alternatives to the traditional scientific criteria of internal and external validity, reliability, and objectivity. Their new criteria included credibility, transferability, dependability, and confirmability. Also, just as quantitative researchers had described ways to establish internal and external validity, Guba and Lincoln discussed ways to establish or judge constructivist evaluations on their new criteria. But they also proposed various forms of authenticity as new criteria, unparalleled in positivist and post positivist paradigms, that were unique to judging the quality of research and evaluation efforts emerging from the constructivist paradigm. The authenticity of a study was established by its fairness (the extent to which it represented different views and value systems associated with the subject of study), and by its ability to raise stakeholders' awareness of issues, to educate them to the views of other stakeholders, and to help them move to action. These ideas, developed at length in their writings, moved evaluators to consider not only other ways to judge evaluation studies, but also to more broadly consider the purposes of their work in, for example, stimulating dialogue and action among stakeholders.

Naturalistic Evaluation. What was Guba and Lincoln's naturalistic evaluation? According to Guba and Lincoln, the major role of evaluation is one of responding to an audience's requirements for information in ways that take into account the different value perspectives of its members. By taking a naturalistic approach to evaluation, the evaluator is studying the program activity in situ, or as it occurs naturally, without constraining, manipulating, or controlling it. Naturalistic inquiry casts the evaluator in the role of a learner and those being studied in the role of informants who *teach* the evaluator. The dominant perspective is that of the informants, because the evaluators learn their perspectives, learn the concepts they use to describe their world, use their definitions of these concepts, learn the folk theory explanations, and translate their world so the evaluator and others can understand it.

The naturalistic approach, and its benefits, is perhaps best articulated by a non-evaluator, Elspeth Huxley. In *The Flame Trees of Thika*, she astutely observed:

> The best way to find things out is not to ask questions at all. If you fire off a question, it is like firing off a gun—bang it goes, and everything takes flight and runs for shelter. But if you sit quite still and pretend not to be looking, all the little facts will come and peck round your feet, situations will venture forth from thickets, and intentions will creep out and sun themselves on a stone; and if you are very patient, you will see and understand a great deal more than a man with a gun does. (1982, p. 272)

The naturalistic evaluator, like Huxley, observes the program and its actions, its participants and staff, in its natural environment and, through observations, documents and records, interviews, and unobtrusive measures, comes to understand and describe the program.

Stakeholder Participation. Guba and Lincoln's approach to evaluation also moved evaluators to consider more participatory approaches. Like Stake, Guba and Lincoln believed evaluation should represent the multiple realities seen by stakeholders, but their models advocated a more active role for stakeholders than did Stake with responsive evaluation. They saw the evaluator's role as one of gathering views and values from a wide variety of stakeholders and, then, with the evaluator in the role of negotiator, working with stakeholders, to inform each other about the program and determine steps toward action. Evaluators were no longer simply measuring, describing, and judging—the first three generations of evaluation noted in their book, *Fourth-Generation Evaluation*. Instead, the fourth generation entailed acting as a negotiator to help stakeholders reach consensus on their diverse views and decide on next steps or priorities. Greene writes, "With this promotion of evaluation as negotiation, Guba and Lincoln explicitly situated evaluation as a value-committed, emancipatory practice in contrast to most prior evaluation theories which either claimed value neutrality or value pluralism" (2008, p. 323).

Writing in 2004, Lincoln and Guba noted how much evaluation had changed since their writings in the 1980s: "[S]ince the last time we wrote in any extended way about evaluation as model, set of practices and discourses, or theoretical perspective, the evaluation world has become far more complex and more sophisticated (as well as more populous)" (2004, p. 226). But they observe

how their work foreshadowed some of today's approaches and writings in participatory evaluation. Stake's responsive evaluation introduced evaluators to more qualitative approaches and to learning the views of stakeholders. Guba and Lincoln's naturalistic and fourth-generation books provided an epistemological foundation for naturalistic inquiry but, unlike Stake, argued for more active roles for stakeholders and less neutral roles for the evaluator, roles in which the evaluator moved to act for the stakeholders. This contribution is reflected in many contemporary participatory models.

Having introduced the reader to some of the founders of participatory evaluation, let us move to describing the many participatory approaches that have emerged today. Current approaches mirror the similarities and differences seen both in Stake's and in Guba and Lincoln's models.

Participatory Evaluation Today: Two Streams and Many Approaches

Many of today's participatory approaches to evaluation have evolved from Stake's and from Guba and Lincoln's writings and approaches from the 1970s and 1980s. In fact, there are so many that we will not be able to summarize them all here. Instead, we will first describe a method for categorizing the numerous models and will then describe some of the better known models to give the reader a sense for the differences and the commonalities across contemporary participative approaches.

But before doing so, we will attempt to define participatory evaluation as the term is used today. In the *Encyclopedia of Evaluation*, Jean King defines participatory evaluation as "an overarching term for any evaluation approach that involves program staff or participants actively in decision making and other activities related to the planning and implementation of evaluation studies" (2005, p. 291). Cousins and Earl, in their important work concerning the theoretical foundations and practice of participatory evaluation, define participatory evaluation as "applied social research that involves a partnership between trained and practice-based decision makers, organization members with program responsibility, or people with a vital interest in the program" (1992, p. 399). These definitions are very broad and reflect what Cullen (2009) and others (Cousins & Whitmore, 1998; O'Sullivan & D'Agostino, 2002) describe as the confusion around the meaning of the term and its overlap with other terms such as collaborative evaluation.[2] Cullen's study of evaluation practice in international development finds that the term has many different meanings to the international

[2]Cousins and Whitmore (1998) and others typically use *collaborative* evaluation as an umbrella term that includes participative evaluation. Rodriguez-Campos (2005), however, has written a book on collaborative evaluation proposing it as a different approach. Cullen (2009) argues that the approach overlaps with practical participatory evaluation. These differences in views reflect the diversity of approaches. Each author, of course, may add nuances, or larger differences, that can be useful to the individual evaluator.

evaluators she surveys and interviews. King, too, acknowledges this problem, but argues that participatory approaches are distinguished by four characteristics:

1. Direct and active participant involvement over time in evaluation planning and implementation. It is stakeholder based or consumer focused and is "a way to democratize the evaluation process."
2. Participant ownership which increases the likelihood that participants will use the results.
3. A professional evaluator provides technical assistance, and serves as a "partner, facilitator, or coach—a teacher or consultant."
4. The evaluation capacity of individuals or an organization, their understanding of the evaluation process, and their actual evaluation skills are likely to increase (King, 2005, p. 291).

Further, King notes two frequent misunderstandings regarding participatory evaluation: (a) the belief that any involvement of program staff or participants, including simply collecting data from them, makes something a participatory evaluation; and (b) the view that any evaluation that uses qualitative methods is participatory. The latter confusion comes partly from Stake's and Guba and Lincoln's pairing of qualitative methods and participatory approaches. In fact, those who favor participative approaches often do see the benefits of qualitative data. But whether one collects qualitative or quantitative data or a mix of various types of data is a methodological decision. It is not a decision concerning the approach to the evaluation—the principles that will guide the planning and conduct of the evaluation. King notes that evaluators collecting qualitative data often interact with stakeholders through interviews or observations to collect that data. But, she emphasizes,

> It is not this direct contact that defines a participatory evaluation. . . . Rather, it is the nature of the relationship between them [the evaluator and stakeholders] that determines whether or not the study is participatory. If the evaluator retains complete control of decision making for the study, regardless of the data-collection methods used or the time spent on site, the study is not participatory. (2005, p. 292)

Categories of Participatory Approaches

As participatory approaches proliferated, evaluators led by Bradley Cousins began to identify characteristics or dimensions to help distinguish among the approaches. Cousins, Donohue, and Bloom (1996) specified three dimensions in which participatory approaches differ:

1. *Control over the evaluation or technical decision-making process.* Does the evaluator maintain sole or primary control? Or, at the other end of the continuum, do the stakeholders have primary control over the conduct of the evaluation, with the evaluator providing technical advice?
2. *Stakeholder selection.* How broad or diverse an array of stakeholders is included in the participation process? Are only primary users, such as the funding source and a few selected recipients or managers, included? Or, does the participation include all legitimate stakeholder groups?

3. *Depth of participation.* In what ways do stakeholders participate in the evaluation? Do they participate in every phase, or is their participation limited to input on a few, non technical issues?

In a widely cited study, Cousins and Whitmore (1998) then used these dimensions to analyze and categorize ten different participative approaches to evaluation and action research.[3] The approaches they reviewed included some well-known participatory evaluation approaches: their own practical participatory evaluation (Cousins & Earl, 1992, 1995), Mark and Shotland's stakeholder-based evaluation (1985), Patton's developmental evaluation (1994, 2010), and Fetterman's empowerment evaluation (1994, 1996, 2001a, 2005), as well as several types of action research. They assessed each approach on three dimensions: control of the evaluation, the selection of stakeholders to be involved, and the depth of participation of those stakeholders. Cousins and Whitmore found variation at each level that helped to compare and contrast the current approaches to participatory or collaborative evaluation that are discussed in the literature. We will make use of their analysis to acquaint the reader with current types of participatory evaluation and their characteristics.

First, Cousins and Whitmore observe that contemporary participative approaches can be narrowed to two primary types: practical participatory evaluation (P-PE) and transformative participatory evaluation (T-PE). These two streams have different histories, purposes, and methods. Practical participatory evaluation, as the name implies, is used for practical reasons limited to the program being evaluated and the organization in which it resides. Specifically, these participatory approaches involve stakeholders in the evaluation to improve the usefulness of the results. As they write, "The core premise of P-PE is that stakeholder participation in evaluation will enhance evaluation relevance, ownership, and thus utilization" (1998, p. 6). Although P-PE was developed primarily in the United States and Canada, T-PE first emerged in the developing world, including Central and South America, India, and Africa and, unlike P-PE, emerged from community and international development and adult education (Fals-Borda & Anisur-Rahman, 1991; Freire, 1982; Hall, 1992). The purpose of transformative evaluation is, in fact, to transform, to empower stakeholders, who have been relatively powerless, by their participation in action research or evaluation. This participation provides them with self-knowledge and skills and an understanding of the power arrangements concerning their program and their locality. Transformative evaluations, while concerned with the evaluation of a particular program, are also intended to bring about social change. Its purposes are broader than the program being evaluated and are explicitly political. These approaches, largely emerging from action research, are designed to change power structures; to empower oppressed people, especially those in rural developing countries; and to reduce poverty. Today, transformative evaluation approaches are also seen in university-based work in the United States (Mertens, 1999, 2001, 2008).

[3]Action research preceded evaluation in having an emphasis on participants or stakeholders conducting the research. In fact, the purpose of most action research is to provide professionals with models and tools for conducting their own research to improve their practice.

Transformative evaluation differs from P-PE not only in its purposes, but also in its methods. To achieve its transformative purposes, transformative evaluators serve more as consultants with the control of the evaluation being in the hands of the stakeholders, particularly those with less power. The evaluation must be guided and driven by these stakeholders so their participation will provide them with greater knowledge, skills, and power. But, delegating responsibilities to groups who are new to research and evaluation can prompt concerns about the validity of the results. Advocates argue that the delegation of responsibility and the skills gained are more important outcomes than the validity of the study, but they also argue that the contributions of stakeholders most familiar with the local context can increase the validity of the evaluation.

Differences in Current Participatory Approaches

Our purposes here are to help the reader distinguish among prominent current participatory approaches. Therefore, we will return to Cousins and Whitmore's analysis of ten different models using their three dimensions: control of the evaluation, stakeholder selection for participation, and depth of participation. In Table 8.1, we summarize their findings on the ten approaches they reviewed. Regarding control of the evaluation, half of the approaches pursue a balance between evaluator and stakeholder control. These include Cousins and Earl's practical participatory evaluation and Patton's developmental evaluation. Only one approach, Mark and Shotland's stakeholder-based evaluation, includes the traditional evaluator-led control of decision making. This approach most resembles Stake's responsive evaluation and, in fact, studies of evaluators show that variants of this approach are the ones most commonly used in participatory evaluations in the United States and Canada (Cousins, Donohue, & Bloom, 1996) and in international, developmental evaluations (Cullen, 2009). Four of the approaches, however, fall at the other end of the continuum, with stakeholders dominating on decisions regarding the evaluation. One of these approaches, not surprisingly, is T-PE, the context of which is primarily developing countries. David Fetterman's empowerment evaluation is the only other approach that tips the control

TABLE 8.1 ***Summary of Characteristics of Ten Forms/Approaches of Systematic Inquiry from Cousins and Whitmore****

Dimension	*Number of Forms/Approaches in Each Category*		
Control of evaluation process	Evaluator	Partnership	Stakeholder
	1	4	5
Stakeholder selection	Limited		Many/All
	4	0	6
Depth of participation	Consultation	Mid-level	Deep Participation
	2	1	7

*Data summarized from Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. In E. Whitmore (Ed.), *Understanding and practicing participatory evaluation.* New Directions for Evaluation, No. 80, 5–23. San Francisco: Jossey-Bass. With permission

of power to the stakeholder; the remaining two are action research approaches rather than evaluation approaches (Carr & Kemmis, 1992 and McTaggart, 1991, with Emancipatory Action Research) or cooperative inquiries (Heron, 1981; Reason, 1994).

The ten approaches studied are similarly diverse in their placement on the continuum of stakeholder selection. Three well-known evaluation approaches—Cousins and Earl's practical participatory evaluation, Fetterman's empowerment evaluation, and Patton's development evaluation—and one action research approach, Argyris and Schöen's participatory action research—limit stakeholder involvement to managers or policymakers or, as Cousins and Whitmore write, they work "in partnership with potential users who have the clout to do something with the evaluation findings or emergent recommendations" (1998, p. 11). Six approaches, however, opt for broad stakeholder involvement, including many diverse groups. These approaches include both the frequently used stakeholder-based evaluation and, of course, T-PE. Finally, most approaches advocate deep involvement of the stakeholder groups that are included in the study, going beyond selecting evaluation questions and interpreting results. Only one approach, stakeholder-based evaluation, limits participation to helping in identifying evaluation purposes and questions and, in the final stages, interpreting results and identifying recommendations. Most approaches involve stakeholders in all phases of the evaluation in one way or another.

In summary, Cousins and Whitmore's review of participatory approaches to inquiry reveals that these approaches differ significantly in how they advise evaluators to manage participation. Some, like Mark and Shotland's stakeholder-based evaluation, keep the control of the evaluation in the hands of the evaluator. However, they are participative in the sense that they involve a range—often a wide range—of stakeholders in considering what questions the evaluation should address and how to interpret results and make recommendations. Others, represented by Cousins and Earl's P-PE, Patton's developmental evaluation, and Participatory Action Research, balance control between the evaluator and the stakeholder. They involve primarily managers and policymakers as stakeholders, but seek intensive involvement of these stakeholders. Other, more transformative approaches, cede control of the study to a wide variety of stakeholder groups who, then, have a great depth of participation.

All these approaches involve stakeholders, but the diversity in the models provides the reader with many options for pursuing a participatory approach and many issues to consider. There is no template for conducting a participatory evaluation. Instead, readers can, and should, choose an approach or dimensions of participation that work most effectively in the context of the program they are evaluating. Transformative evaluations arose in developing countries and in countries where program participants were often marginalized, oppressed, and living in serious poverty. Social justice was, and remains, a major consideration in evaluation in such settings.[4] Cousins and Earl argue that their participatory approach of actively seeking participation and sharing responsibility but limiting participation generally to managers, staff, or policymakers, can be most appropriate when the purpose of the evaluation is formative, that is, making decisions for program improvement.

[4]Social justice concerns exist in developed countries as well. Government reactions to Hurricane Katrina raised U.S. citizens' awareness of this issue.

In such a case, involving stakeholders who can make decisions is important in order to meet their information needs, gain their trust, and increase the usefulness and actual use of the evaluation. In contrast, stakeholder-based evaluation, in which the evaluator maintains control but seeks input from many different stakeholders on beginning and ending concerns in the evaluation, can be most appropriate for summative decisions, such as deciding whether to continue a program. Such a major, political decision must include input from many different groups concerning the focus of the evaluation and the interpretation of the results. However, it may be useful for the evaluator to maintain control of the technical decisions in the study to ensure that the methodological choices and results are not questioned in the highly political context of many summative evaluations. Table 8.2 lists some participatory evaluation approaches and the contexts in which each approach might work best.

TABLE 8.2 ***Participatory Evaluation Approaches and Contexts: What to Use When***

Approach	*Key Elements*	*Context in Which It Fits*
Practical Participatory Evaluation (P-PE) (Cousins and Earl)	Balanced control; Stakeholders—managers, staff, policymakers; Much involvement	Formative decisions
Developmental Eval. (Patton)	Balanced control; Stakeholders—team members, policymakers, managers; Each team member involved	Development activities; Changing environment
Transformative Participatory Eval. (T-PE)	Stakeholder control; Many stakeholders; Emphasis on program participants; Much involvement	Participants are oppressed; Social justice is concern; Participant empowerment goal
Stakeholder-based Evaluation (Stake; Mark & Shotland)	Evaluator control; Many stakeholders; Limited involvement	Summative; Tech expertise, validity imp; Input on questions and results help direct study improve use
Empowerment Evaluation (Fetterman & Wandersman)	Stakeholder control; Stakeholders—managers, staff; Much involvement	Need for empowerment and capacity building of staff; Need for building internal mechanism for self-monitoring and program improvement
Deliberative Democratic Eval. (House & Howe)*	Evaluator control; Many stakeholders; Limited involvement	Need for dialogue among stakeholders; Power differentials among participants require evaluator to moderate

*House and Howe's deliberative democratic evaluation approach post dated Cousins and Whitmore's analysis. It is rated by the authors of this textbook based on House and Howe's writings and Cousins and Whitmore's dimensions.

Some Specific Contemporary Approaches

Practical Participatory Evaluation (P-PE)

Bradley Cousins has been a leader in writing about participatory evaluation, studying its various approaches, collecting empirical data on evaluation practices and use, and reviewing research from many related fields such as adult learning, organizational learning, knowledge construction, and evaluation theory and practice. Grounded in this literature and research, his writing on his P-PE approach is, as he and Earl write, "light on prescription and comparatively heavy on justification" (1992, p. 397). He continues that theme more recently in describing the means by which various collaborative approaches, including his P-PE, Fetterman's empowerment evaluation, and Patton's developmental approach, set program standards (Cousins & Shulha, 2008), noting that the P-PE approach to setting standards should be an emergent approach adapting to the context while working with primary stakeholders. In other words, it is difficult to describe P-PE in depth because Cousins acknowledges, even emphasizes, the need for adaptation to context.

Cousins and Earl (1992) first articulate P-PE as a method to increase use, explicitly focusing on evaluation's decision-oriented purpose and building on the research over the years concerning its use in various fields. Their first paper is largely a review of 26 different studies on the relationship between evaluator-practitioner linkages and evaluation use. From these studies, and other research and theories on organizational learning (Argyris & Schöen, 1978; Senge, 1990), Cousins and Earl develop a P-PE approach that builds on the following evidence from research:

- Use of evaluation results is enhanced by communication, contact, and collaboration between evaluators and primary stakeholders; that is, those who are most interested in results and in a position to use them.
- Evaluators should focus less on the use of a particular study and more on learning about the organization and its context in order to provide useful information in an ongoing fashion (Weiss & Bucuvalas, 1980).
- Knowledge or information is "socially constructed," meaning knowledge is based on one's images or interpretations of reality, not the precise details of reality (Bandura, 1977, 1986).
- Like individuals, organizations develop their own views of reality among employees and within the organizational culture based on shared images and mental models of the organization (Argyris & Schöen, 1978; Senge, 1990).
- By establishing linkages with those in the organization, spending time in the organization to learn its images and culture, and involving primary stakeholders intensely, as partners, in doing the evaluation, the evaluator will increase the chances that the results will be used. More importantly, the primary stakeholders who work on the study can continue evaluation activities or evaluative ways of thinking.

- Involving these primary stakeholders will enhance organizational learning by changing images and views and even ways of establishing those images and views, such as by questioning core assumptions and collecting data or information to determine what works.

Cousins and Earl's (1992, 1995) P-PE approach involves evaluators working closely with primary stakeholders, program managers, and staff to jointly conduct an evaluation. The approach includes training key organizational personnel in the technical skills of evaluation, while they are working in partnership with evaluators. This form of capacity building in organizations is intended to directly enhance organizational learning and to prepare key organization members to take on evaluation coordination in both continuing and new projects. The evaluator may then move into the role of consultant on technical issues and on tasks related to future evaluation activities. It is best suited, in their view, to formative evaluations that help to inform and improve program implementation.

Differences from Traditional Stakeholder-Based Evaluation. Cousins and Earl contrast their approach to participatory evaluation with traditional stakeholder-based evaluation:

1. In traditional stakeholder-based evaluation, the evaluator controls decisions regarding the evaluation. In P-PE, the evaluator works in partnership with the primary stakeholders and shares in decisions regarding the evaluation. In this way, primary stakeholders gain skills and a sense of ownership about the evaluation, and the evaluator learns more about the program to be evaluated, its issues and politics, and the organizational culture. The evaluator, as the coordinator, is responsible for technical support, training, and quality control, but conducting the study is a joint responsibility.

2. In traditional stakeholder-based evaluation, the evaluator works with representatives of many stakeholder groups to portray all their views and perspectives. In P-PE, the evaluator works with a more limited group of stakeholders, those who can make changes in the program. This group was first identified as "primary stakeholders" by Alkin (1991), who noted that he would prefer to work with stakeholders who are interested in the evaluation and have the power to do something with the results than with many stakeholders who may not be interested in the evaluation or may lack the power to use it. Unlike Stake, who saw his role as portraying the multiple realities seen by an array of stakeholders, Cousins and Earl, and Alkin, select stakeholders for their potential to use the study.

3. In a traditional stakeholder-based evaluation, the many stakeholders provide relatively limited input to the study, typically helping to define the purpose of the study and the questions it will address and to provide their perspectives on the program. In P-PE, the smaller group of primary stakeholders is thoroughly involved in every stage of the evaluation. This in-depth involvement is designed to increase their sense of ownership and their understanding of the

study as well as their ability to use evaluative means of thinking and methods in the future. Stakeholders' partial involvement in only a few stages would not achieve this goal.

P-PE is designed to improve use. Its primary purpose is practical, not political as in T-PE. Cousins and Earl's approach is not designed to empower participants or to change power distributions. It is, however, designed to encourage organizational learning and change. Although the immediate goal is practical, increasing the usefulness and actual use of the current evaluation, the long-term goal is capacity building (giving existing staff or managers skills in evaluation) and creating a learning organization that makes use of evaluation information for planning and improvement. Cousins and Earl acknowledge that the approach would not be particularly useful for summative evaluations. Objectivity would be a primary concern and the participating role of the primary stakeholders (people responsible for the program) in conducting the evaluation would be problematic in such circumstances. Also, summative evaluations often require involvement of many stakeholder groups to deal with the more political decisions to be made.

Developmental Evaluation

Michael Patton is well known for his utilization-focused evaluation (UFE) approach, with his current book in its fourth edition (Patton, 2008a). However, he has more recently written about developmental evaluation, which he sees as a type, or an option, in a utilization-focused process. We will describe developmental evaluation here because its focus is participative, though it differs substantially from the other models. The approach shows how participative approaches and even definitions of evaluators' roles develop and change as our experience with organizations, participation, and evaluation change and as the nature of an organization's work changes.

The Evaluator Moves into New Roles. Cousins and Whitmore used stakeholder participation to increase the use of formative evaluations and to increase learning in the organization over the long run. With developmental evaluation, Patton moves the evaluator into the planning and implementation of programs or other developmental activities of the organization. The stakeholders aren't members of an evaluation team. Instead, the developmental evaluator is a member of the program team. The evaluator isn't training other team members in evaluation skills; instead, those are the skills the evaluator brings to the team just as other members bring their own areas of expertise. Finally, developmental evaluation doesn't evaluate a particular thing. It uses evaluative modes of thinking and techniques to help in the constant, ongoing, changing development process and growth of an organization. In his work with organizations on developmental issues, Patton has noted and emphasized that these environments differ from those in which traditional program evaluations are conducted. Environments for development are characterized by complexity, turbulence, and the dynamic, nonlinear nature of the

environments (Patton, 2009, 2010). Thus, he argues, new approaches and new roles for evaluators are needed.

This is turning evaluation on its head! Developmental evaluation is an entirely new type of participatory approach. The team isn't the stakeholder and the evaluator. It is a group of people with many different areas of expertise necessary to plan and guide the organization. Patton writes that "developmental evaluation isn't a [evaluation] model. It's a relationship founded on a shared purpose: development" (1994, p. 313). He describes it as a "long-term partnering relationship with clients who are, themselves, engaged in on-going program development" (1994, p. 312). And, what is development? It is the ongoing work of an organization or even a group that is exploring an issue and some solutions to that issue. Patton gives examples of a group working on community leadership in rural Minnesota, another group working with St. Paul Public Schools to build and support multicultural education in the district, and still another group working on a 20-year initiative to improve community health in two inner-city neighborhoods. Development projects differ from traditional evaluations in that there are no clear goals, which is seen as limiting the development, and no established timeframe. The evaluator isn't working to develop an evaluation report to provide to an external funder at some given point in time. Instead, the team is constantly tinkering to deal with changes—changes in what they know, in what participants need, and in the context of the community. What does the evaluator bring to the table? Patton argues that experienced evaluators, during their years of practice, have gained more than evaluation expertise. They have gained skills in the logic of program development: "We know a lot about patterns of effectiveness. . . . That knowledge makes us valuable partners in the design process" (Patton, 2005b, p. 16). The evaluator serves as a team member who raises evaluation-type questions using evaluative logic and knowledge of research, and supports using data and logic in the decision-making process.

In developmental evaluation, evaluators are explicitly part of the program design team. Thus, the evaluator loses the external, independent role that can be valuable in summative decisions. But developmental evaluation is not intended for summative decisions, nor is it for formative decisions. It is for development. Acknowledging that developmental evaluation is akin to organizational development (OD), Patton argues that evaluators need to share their skills: "When evaluation theorists caution against crossing the line from rendering judgment to offering advice, they may underestimate the valuable role evaluators can play in design and program improvement" (Patton, 2005b, p. 16). (See also Reichardt [1994] advocating that evaluators play a greater role in program development rather than focusing on summative evaluations as a way to have a greater impact on organization and program improvement.[5])

[5]Patton's book on developmental evaluation is being written at the same time that this textbook is being updated. He has shared the first chapter with us, but we have relied on existing publications here. We encourage readers interested in this approach to read his new book, cited in suggested readings at the end of the chapter.

Empowerment Evaluation

The two contemporary participatory approaches reviewed previously fall in Cousins's and Whitmore's practical stream. The next two approaches we will describe, empowerment evaluation and democratic evaluations, are from the transformative stream; their methods have political purposes or foundations. The political purposes of empowerment evaluation are to empower their stakeholders and to help them achieve self-determination in evaluation. David Fetterman introduced empowerment evaluation in his presidential address to the American Evaluation Association in 1993 (Fetterman, 1994). He and his colleague, Abraham Wandersman, have since published three books on the approach, providing many tools and examples as well as discussion of the merits of the approach and its foundations (Fetterman, 2001a; Fetterman, Kaftarian, & Wandersman, 1996; Fetterman & Wandersman, 2005). Unlike P-PE, Fetterman sees empowerment evaluation as emerging for explicitly political purposes—to empower stakeholders through self-determination. Therefore, it would be considered in the transformative stream. However, both Cousins and Whitmore (1998), and Fetterman and Wandersman themselves (2007), raise questions about the extent to which the approach's purposes and foundation are transformative. Cousins and Whitmore, in their analysis of ten collaborative approaches, viewed empowerment evaluation as "enigmatic" and differing from other transformative approaches in practice, because empowerment evaluators tend to work more closely with the evaluation and with a more limited group of stakeholders—generally program staff and managers—than do evaluators in most transformative approaches. In 2007, Fetterman and Wandersman argued that the empowerment approach can be practical, to facilitate use and program improvement, or transformative, to empower stakeholders, depending on the setting and the purpose.

Definitions of Empowerment Evaluation. Fetterman's original definition of empowerment evaluation focused on the empowerment aim: "Empowerment evaluation is the use of evaluation concepts and techniques to foster self-determination. The focus is on helping people help themselves" (1994, p. 1). In his 2001 book, he essentially maintained that definition, but modified it to include using evaluation findings and serving purposes of improvement, writing that empowerment evaluation involved "the use of evaluation concepts, techniques, and findings to foster improvement and self-determination" (2001a, p. 3). In 2005, Fetterman and Wandersman's definition emphasized the more practical purpose of organizational learning as they defined empowerment evaluation as "an approach that aims to increase the probability of program success by (1) providing stakeholders with tools for assessing the planning, implementation, and self-evaluation of their program, and (2) mainstreaming evaluation as part of the planning and management of the program/organization" (2005, p. 28). Fetterman and Wandersman (2007) see this most recent definition as a refinement of the earlier definition based on extensive practice with the approach. It also reflects a movement away from a singular emphasis on self-determination and empowerment, or transformation, to an emphasis on capacity building and establishing evaluation systems within the organization to guide it for program improvement.

Differences from Traditional Stakeholder-Based Evaluation. Continuing our contrast with traditional stakeholder-based participatory evaluation, as we note in Figure 8.2, empowerment evaluation differs from traditional stakeholder evaluation on all three dimensions:

1. Rather than the evaluator having control of the decisions concerning the evaluation, the designated stakeholder group or groups to be empowered have control. The evaluator serves as a coach or guide, but stakeholders may overrule the evaluator.

2. Typically, in empowerment evaluation, not all stakeholder groups are involved in the evaluation as they are in the traditional stakeholder evaluation. Instead, the groups to be empowered are the selected participants. Most often, because empowerment evaluation is setting up a system for self-monitoring of organizational performance, those groups (the *self*) are managers and staff who operate the program, though program recipients can be involved.

3. Unlike traditional stakeholder evaluation, but like P-PE, the smaller number of stakeholders selected for participation is intensely involved in carrying out the evaluation, more so than in Cousins and Earl's P-PE, where the evaluator and stakeholder roles are balanced. The stakeholders are making the decisions, with guidance from the evaluator at all phases. Empowerment evaluation provides many tools to assist stakeholders in these tasks, from "taking stock" to establish a baseline for future comparisons of program performance to strategies for planning for the future (Getting To Outcomes).

What then is empowerment evaluation? In his initial presentation, Fetterman (1994) revealed several different roles an evaluator plays in an empowerment evaluation. These included the following:

1. Training: "Evaluators teach people to conduct their own evaluations and thus become more self-sufficient" (p. 3).
2. Facilitation: "Evaluators can serve as coaches or facilitators to help others conduct their evaluation" (p. 4).
3. Advocacy: "Evaluators may even serve as direct advocates—helping to empower groups through evaluation." . . . "(p. 6) "Advocate evaluators write in public forums to change public opinion, embarrass power brokers, and provide relevant information at opportune moments in the policy decision making forum" (p. 7).

But a key issue is that the stakeholders are taking the lead. The evaluator is only facilitating. Fetterman illustrates the evaluator's role with this analogy:

> Empowerment evaluation helps to transform the potential energy of a community into kinetic energy. However, they [stakeholders] are the source of that energy. Standing at a baseball diamond, it is not the bat that drives the run home; it is the player. The bat, like the empowerment evaluator, is only an instrument used to transform that energy. (2007, p. 182)

Empowerment Evaluation Methods. Fetterman and Wandersman, like the bat, have developed particular tools to assist stakeholders in undertaking an evaluation. These are the 3-step approach and the 10-step Getting To Outcomes (GTO) approach. The 3-step approach illustrates the basic stages of empowerment evaluation:

1. Establish the mission or purpose of the program to be evaluated.
2. Take stock of the current state of affairs of the program in regard to the purpose. This stage is considered the baseline to be used for future comparison to assess program progress.
3. Plan for the future to achieve the articulated purpose. These plans represent the group's intervention, which will be evaluated (Fetterman & Wandersman, 2007, p. 187).

In a second, taking-stock session, after program changes have been implemented, the stakeholders then compare the new, current state of affairs with the previous baseline to judge program success. As Fetterman and Wandersman note, the bottom line is, "Are the groups accomplishing their objectives and achieving desired outcomes?" (2007, p. 187).

The following 10-step approach lays out a more detailed intervention using questions that stakeholders should ask to consider how to improve their program, and activities (in parentheses) that the stakeholders should undertake to answer these questions.

1. What are the needs and resources in your organization, school, community, or state? (needs assessment; resource assessment)
2. What are the goals, target population, and desired outcomes (objectives) for your school/community/state? (goal setting)
3. How does the intervention incorporate knowledge of science and best practices in this area? (science and best practices)
4. How does the intervention fit with other programs already being offered? (collaboration; cultural competence)
5. What capacities do you need to put this intervention into place with quality? (capacity building)
6. How will this intervention be carried out? (planning)
7. How will the quality of implementation be assessed? (process evaluation)
8. How well did the intervention work? (impact and outcome assessment)
9. How will continuous quality improvement strategies be incorporated? (total quality management; continuous quality improvement)
10. If the intervention is (or components are) successful, how will the intervention be sustained (sustainability and institutionalization) (adapted from Wandersman, Imm, Chiman, & Kaftarian [2000] by Fetterman & Wandersman [2007] p. 188).

These questions demonstrate that as empowerment evaluation has evolved, being used in many different settings, the focus has become more on helping stakeholders, typically program managers and staff, to plan new programs that

will achieve their goals and/or to improve existing programs. Answering these questions will empower program staff and managers by leading them into planning and management exercises. (Wandersman et al. [2000] identified the first six questions as planning questions and the last four as evaluation questions.) Unlike other means of participatory or transformative evaluation, empowerment evaluation has become a whole system, an intervention, to improve the organization. This expansion may speak to its popularity. Fetterman and Wandersman provide numerous examples of empowerment evaluation applications in real-world settings. These include its use by the Stanford School of Medicine in accreditation activities, by 18 Indian tribes in California for planning their Tribal Digital Village, and by schools in distress in rural Arkansas to raise achievement scores (Fetterman & Wandersman, 2005; Fetterman & Wandersman, 2007).

Controversies over Empowerment Evaluation. Empowerment evaluation, however, has encountered more controversy and criticism than any other evaluation approach. We will summarize a few of those criticisms, some broad and some focused. First, it has been criticized for conceptual ambiguity. Some have argued that it is not distinct from other forms of participative or collaborative evaluation and that examples of empowerment evaluation frequently depart dramatically from its principles (Patton, 1997b; Patton, 2005a; Cousins, 2005; Miller & Campbell, 2006). Cousins (2005), when asked to write a summary chapter on the cases presented in Fetterman & Wandersman (2005), analyzed the cases using his dimensions for analyzing participatory evaluation approaches and found a great deal of difference across the cases presented, even on the core dimension for empowerment evaluation—stakeholder control.

Miller and Campbell (2006) conducted a comprehensive study of 47 published articles that claimed to use empowerment evaluation. Just as Cullen (2009) found mixed definitions and applications of participatory evaluation, Miller and Campbell found several different trends in the implementation of empowerment evaluation. Specifically, they observed three rather different roles that evaluators took in these evaluations:

(a) A Socratic coaching role, that is, engaging in question and answer sessions with stakeholders (32% of cases, which adhered most closely to the empowerment principles);
(b) Providing structured guidance in which the evaluator designed a set of steps for stakeholders to implement. This style often involved teaching the stakeholders to use a set of empowerment templates to conduct the evaluation (36% of cases);
(c) More traditional participatory evaluation in which the evaluator sought the input of stakeholders but guided and conducted the study himself (30% of cases).

Like Cullen in studying participatory evaluation more broadly, Miller and Campbell concluded that empowerment evaluation is practiced in a wide variety

of ways, most of which do not capture it's core emphasis on self-determination and empowerment. They write:

> The larger picture that emerges from these data suggests that although many projects get labeled (and relabeled) as empowerment evaluations, frequently, these evaluations do not embody the core principles that are supposed to undergird empowerment evaluation practice.
>
> Across all modes of empowerment evaluation practice, the core principle of using community knowledge was nearly ubiquitous, but the principles of democracy, social justice, and using evidence-based practice were particularly infrequent (Miller & Campbell, 2006, p. 314).

Others have been concerned with Fetterman's emphasis on advocacy. Stufflebeam notes that "while his commitment to '. . . helping people help themselves' is a worthy goal, it is not a fundamental goal of evaluation. Surely this is a valuable role that evaluators and all citizens should play, but it is not evaluation" (1994, p. 323) and reminds us that evaluation concerns investigating or judging the merit or worth of an object, not empowering others to do so. More specifically, he writes:

> The approach advocated by Dr. Fetterman gives over authority to the client/interest group to choose criteria, collect data, and write/edit and disseminate reports, all in the name of self-evaluation for empowerment. The client/interest group seems to be given license to tell some preferred story, obtain the evaluator's assistance in getting the story across to constituents or others, possibly project the illusion that the evaluation was done or endorsed by an outside expert, and remain immune from a metaevaluation against the standards of the evaluation field. (Stufflebeam, 1994, p. 324)

The purposes of both evaluation and empowerment evaluation have broadened somewhat since Stufflebeam's comments, but the potential for bias remains a particular problem cited by critics of the empowerment approach. One element of bias cited by critics concerns the manner in which the approach is promulgated. Since Fetterman's unabashed urging of evaluators to "spread the word" about empowerment evaluation, critics have seen it as too concerned with values and advocacy and not sufficiently with reason. Other presidents of the American Evaluation Association, Lee Sechrest (1997) and Nick Smith (2007), voice what is perhaps the major concern regarding empowerment evaluation: the authors' active role as advocates of the approach.[6] Sechrest and Smith argue that empowerment evaluation often appears to be more of an ideology, a set of beliefs, than an evaluation approach. Their advocacy of the empowerment approach, the generality of the cases, and the occasional overstatements in presenting cases (see Patton [2005a])

[6]Smith notes that empowerment evaluation advocates are not alone in acting as ideologues, arguing that those who advocate randomized control trials (RCTs) also fall into this category.

have caused empowerment evaluation to receive more criticism than other approaches that often present fewer cases, but with more specifics and self-criticism.[7] Smith (2004, 2007) notes that in a political world overwhelmed with ideological arguments on everything from stem cells to health policies to climate change (issues on which data exist), we need evaluation grounded in reason, not ideology, to maintain evaluators' credibility in judging the merit and worth of programs and policies. Patton notes the problem of bias in Fetterman and Wandersman's own presentation of cases of success (Patton, 2005a). He notes that often the evidence is ambiguous and may confuse the role of the evaluation in a particular case with other factors that contributed to the program's success. Similarly, Cousins, in his review of the cases presented in Fetterman and Wandersman's book (2005), regrets that the chapters are not so much cases as they are "reflective narratives or essays" and calls for clearer presentation of empowerment cases to learn more about how such evaluations are actually approached and their effectiveness (2005, p. 203).

A second concern with bias is, of course, that which stakeholders may introduce in evaluating their own program.[8] Although Fetterman (2007) writes that he has seen stakeholders be very critical of their own work because they want to improve to achieve success, that is not the experience of many evaluators. Often stakeholders are most concerned with providing evidence of success to funding sources and want the evaluation to show that success. (See, for example, Fitzpatrick and Bledsoe, [2007]) Regarding the potential problem of bias among stakeholders in conducting evaluations, empowerment evaluation may help stakeholders improve their programs if the evaluator can train them to look at their actions dispassionately and/or the organizational culture lends itself to questioning and data-based decision making. But in today's funding and performance monitoring environment, training managers and staff to look at their actions dispassionately and to report the results accordingly can be a challenge.

Summary. In conclusion, empowerment evaluation has promoted much practice as well as much controversy and discussion in the field of evaluation. We conclude that it has changed from emphasizing empowerment and self-determination, primarily of program participants, to emphasizing capacity building among program staff, thereby empowering them to use evaluation methods to bring about program improvement. In this latter iteration, empowerment evaluation is less controversial and more like other approaches to evaluation that are encouraging stakeholder participation and

[7]For an interesting insight into the vigorous dialogue among evaluation leaders on the empowerment approach, see volume 26(3) of the *American Journal of Evaluation* (2005) for, first, two reviews of Fetterman and Wandersman's newest book, *Empowerment Evaluation Principles in Practice*, by Michael Scriven and Michael Patton; then, two separate responses to these reviews by David Fetterman and Abraham Wandersman; and finally, two separate rejoinders to Fetterman's and Wandersman's responses by Scriven and Patton.

[8]This concern is not exclusive to empowerment evaluation, but is a concern with many participatory models. The more power that stakeholders have in the evaluation, however, the greater the concern that bias, intended or not, may enter into the analysis.

emphasizing the role of evaluators' in building organizations' internal evaluation capacity and mainstreaming evaluation. This was the theme of our co-author James Sanders's AEA presidency in 2001. (See Sanders [2002] for his presidential address.) Empowerment evaluation has mellowed with age. Its emphasis on self-determination seems to have faded in recent books, but its power to build evaluation capacity and mainstream it as a way for doing business has increased. In an AJE issue on evaluation in the twenty-first century, Fetterman emphasizes this view:

> The future of evaluation will be characterized by critical and collaborative relationships. Evaluation will be a collaboration. Citizens will come to the table with a basic knowledge of evaluation as part of their democratic literacy skills. Funders, program staff members, and participants will acquire the capacity to monitor and assess critical aspects of their own performance. They will still need evaluators or critical friends but in a different way and at a much higher level of capacity than evaluators are currently engaged. Data will be routinely used to inform decision-making. Evaluation will be institutionalized as part of program planning and management. (Fetterman, 2001b, p. 381)

Democratically Oriented Approaches to Evaluation

The two transformative, participatory evaluation approaches that are best known in the United States and the United Kingdom build on the values of democracy. First, a word about values: As philosophers of science have recognized for decades, science is not value free. Studies of the practice of science show the interaction of world views, personal views, and scientific research (Godfrey-Smith, 2003; Kuhn, 1962). Similarly, the practice and methodologies of evaluation are not value free. Although evaluators have training that helps them become more aware of their values and the possible influence of those values on their work, evaluations are no more value free than other fields of research. To illustrate, some of our values concern our attempts to collect, analyze, and interpret information or data in an objective way. Many evaluators value objectivity and neutrality. But, as Ross Conner, a well-known evaluator, recently said in an interview, he tends to put the word "objective" in quotes, knowing that he is influenced in interpreting data by his own life experiences (Christie & Conner, 2005, p. 371). The decisions he makes about the purpose of an evaluation study, what data to collect, and how to analyze and interpret it are different from, for example, an evaluator from another country or another continent whose training and world view of what is important or valid are likely to be different. Stake recognized the multiple views of reality and developed his approach with the intention of portraying those multiple views. MacDonald (1974, 1976), House (1980), and House and Howe (1999) chose to ground their approaches to evaluation on values concerned with democracy. Since one value of democracy is equality, their approaches were concerned, though in different ways, with involving many stakeholders and giving stakeholders other than the funders of the evaluation a role. Their approaches were also grounded in social justice and empowerment.

Democratic Evaluation. MacDonald's democratic evaluation is historic in addressing some of the contemporary concerns about the role of the evaluator vis-à-vis society and citizens in the 1970s. MacDonald identified three types of evaluators: the bureaucratic evaluator who served government agencies; the autocratic evaluator who, while external to these agencies and more independent in owning the work of the evaluation and possibly publishing it, was also hired by government agencies; and MacDonald's model, the democratic evaluator who was not serving government agencies, but, instead, was conducting evaluations to serve the public's right to know. Concerned with evaluating educational programs in the United Kingdom, he wrote:

> Democratic evaluation is an information service to the whole community about the characteristics of an educational program. . . . The democratic evaluator recognizes value pluralism and seeks to represent a range of interests in his issues formulation. The basic value is an informed citizenry, and the evaluator acts as a broker in exchanges of information between groups who want knowledge of each other. (MacDonald, 1974, p. 45)

Some of the key principles of MacDonald's approach were to broaden evaluation to serve interests beyond those of the immediate policymakers who typically contracted evaluations. Although the evaluator was primarily in control of the study, he was concerned with seeking input from other stakeholders about the purposes of the evaluation, rather than simply imposing his own purposes or those of the sponsor. To serve the public's right to know, MacDonald was also an early advocate for presenting results in ways that were comprehensible to the general public.[9] In addition, he was concerned with the ownership of the data collected in an evaluation and believed that those from whom data were collected had an ownership in their portion of the data. For example, he sought feedback and permission from those interviewed on his use of their comments. Although MacDonald's approach is not in active use today, it is important for its early introduction of democratic values and other issues such as ownership of data and consideration of wider audiences for evaluation.

Deliberative Democratic Evaluation. More recently, House and Howe (1999, 2000) proposed their deliberative democratic evaluation approach. Although House has written for many years about the need for evaluators to consider their role in helping societies achieve social justice (House, 1980, 1993), his collaboration with Kenneth Howe, an expert in philosophy and education, was his first attempt to develop an approach to evaluation. They note that there are other participative approaches that do much of what they suggest, but the impetus for what they call their framework was to emphasize the principle of democracy and its implications for conducting evaluation.

[9]Presenting results in ways that are accessible and comprehensible to program participants and the public are now part of the Guiding Principles of the American Evaluation Association, but this was a new concept in the 1970s.

Like MacDonald, their framework is built on democratic values, but evaluation had changed and grown over the 25 years since MacDonald's democratic evaluation approach, and their concerns were not with ownership of data or the importance of considering the public as a client for evaluation. Instead, their concerns were with social justice and equality. Democratic values prompted House and Howe to want to involve many stakeholders in the evaluation. By 1999, that was no longer new. Many models of participatory evaluation existed. What was new was their recognition that all stakeholder groups do not have equal power or equal experiences. Disenfranchised groups, often participants served by social programs, do not have the experience or expertise with participation that other stakeholders (policymakers, managers, program staff) have. Placed in a group of people who implement a program or program managers, participants may be reluctant to articulate their needs or concerns for a variety of reasons. As House and Howe note, "Proper deliberation cannot be simply a free-for-all among stakeholders. If it is, the powerful stakeholders win" (1999, p. 94). Thus, they argue, to make the evaluation process fully democratic, the evaluator needs to work to ensure that those groups who traditionally have less power are able to participate in the process.

Deliberative Democratic Principles. To achieve its goals, deliberative democratic evaluation is based on three principles:

- Inclusion of all legitimate, relevant interests.
- Using dialogue to determine the real interests of each stakeholder group.
- Using deliberation to guide stakeholders in a discussion of the merits of different options and for the evaluator to draw conclusions.

Unlike some other participative models (see Figure 8.2), the emphasis in the deliberative democratic model is to include all legitimate stakeholders. As House and Howe write, "The most basic tenet of democracy is that all those who have legitimate, relevant interests should be included in decisions that affect those interests" (1999, p. 5). These stakeholder groups are not as involved as they are in other models. They do not, for example, perform technical tasks. The goal of the deliberative democratic approach is to include all stakeholder groups, not to empower them or to build evaluation competency, but so the evaluator can learn about the real interests and needs of all legitimate stakeholders and can use that information in a deliberative way to reach evaluative conclusions. House does not object to active stakeholder participation, but notes that it generally is not feasible to have all stakeholder groups participate actively. Therefore, those who are active are more likely to be those with the power, resources, and time to participate. House writes, "although I favor actively involving stakeholders under the right conditions, I do not favor actively involving a few stakeholders at the expense of the others" (2003, p. 54). The democratic principles of social equity and equality are addressed by giving priority to inclusion and sacrificing in-depth participation.

Use of Dialogue. The deliberative democratic approach emphasizes dialogue as a means for communicating—sharing views and perspectives. Ryan and DeStefano (2000b) and others (Karlsson, 1998; Preskill & Torres, 1998; Schwandt, 1999) have also written about the important role of dialogue in evaluation practice. Evaluators must engage in frequent deliberative dialogue with stakeholders. That dialogue can take a number of different forms, from simple conversation to collective inquiry, instruction, debate, and so on. During the early stages of an evaluation, dialogue is necessary to overcome the difficulties stakeholders may have in recognizing or articulating their interests, or that evaluators may have in recognizing or understanding their interests. House and Howe are among the few evaluators to note this difficulty, observing, "Individuals and groups are not always able to determine their own interests when left to their own devices. They can be misled by the media or interest groups 'spinning' evidence, or by not having or exercising opportunities to obtain information" (1999, p. 99). So, by engaging in dialogue with the evaluator and other stakeholders, they can better identify their real interests. (See Ryan and DeStefano [2000a] for a more extended discussion of the nature and types of dialogue.)

Deliberation, then, is the last stage, though dialogue and deliberation may overlap. House and Howe emphasize that deliberation is "fundamentally a cognitive process, grounded in reason, evidence, and principles of valid argument, an important subset of which are the methodological canons of evaluation" (1999, p. 101). The evaluator plays a key role in guiding the deliberative process to assist stakeholders in using reasoning and evidence to reach conclusions. So, in another critical way, deliberative democracy differs from other participatory models in that the evaluator plays a key defining role in guiding dialogue and deliberation and in reaching conclusions. He or she does not share critical decisions with the stakeholders. Instead, on the dimension of controlling decision making, the deliberative democratic evaluator is similar to Stake in responsive evaluation or to evaluators using stakeholder-based evaluation approaches. The evaluator leads the deliberative process, but controls for stakeholder bias and other sources of bias. The deliberative process is not intended to help stakeholders reach consensus, but, rather, to inform the evaluation, to help the evaluator learn of the reactions and views of different stakeholder groups and thus to reach appropriate conclusions about merit or worth. Observing the independent role of the evaluator in this approach, Stufflebeam writes approvingly:

> A key advantage over some other advocacy approaches is that the deliberative democratic evaluator expressly reserves the right to rule out inputs that are considered incorrect or unethical. The evaluator is open to all stakeholders' views, carefully considers them, but then renders as defensible a judgment of the program as possible. He or she does not leave the responsibility for reaching a defensible final assessment to a majority vote of stakeholders—some of whom are sure to have conflicts of interest and be uninformed. In rendering a final judgment, the evaluator ensures closure. (2001b, p. 76)

Critics of the model are concerned, however, with the reverse. Some wonder if the promotion of democracy should be the main goal of evaluation (Stake, 2000b) or if the evaluator's values and goals can or should dominate

(Kushner, 2000; Arens & Schwandt, 2000). Transformative evaluations, of course, would go further in trying to bring about political change. House and Howe choose to remain with a value, democracy, that is widely accepted in the United States.

House and Howe note that their approach is an ideal one, and unlikely to be fully implemented in an individual study though its principles are revealed and discussed in several evaluations in an issue of *New Directions for Evaluation* (Ryan & DeStefano, 2000a). They do not see the approach as a guide for conducting an evaluation, but rather as a mid-level theory about evaluation practice. But they note, "Good practice is eclectic and informed by theory, not totally derived from it" (House & Howe, 1999, p. 112). Their approach reminds us to consider inclusion and the democratic principles they represent, but to be aware of power differences that can inhibit participation of disenfranchised groups and of our role in helping those disenfranchised stakeholders to recognize and articulate their own real needs. Finally, their emphasis on deliberation—taking the time to deliberate, to use reasoning to carefully examine all evidence, and to seek input from others—is something evaluators often neglect in the hurried last stages of an evaluation. Reminding us to plan time for deliberation during that last stage is another important feature of their model.

House and Howe also inspire us to consider what evaluation could be:

> If we look beyond the conduct of individual studies by individual evaluators, we can see the outlines of evaluation as an influential societal institution, one that can be vital to the realization of democratic societies. Amid the claims and counterclaims of the mass media, amid public relations and advertising, amid the legions of those in our society who represent particular interests for pay, evaluation can be an institution that stands apart, reliable in the accuracy and integrity of its claims. But it needs a set of explicit democratic principles to guide its practice and test its intuitions. (2000, p. 4)

Looking Back

So we see the multiplicity of participative approaches to evaluation. They have proliferated from Stake's early attention to the multiple views of stakeholders and responsiveness to today's participatory models designed to increase use or transform participants, organizations, or society. They vary in the amount of control evaluators retain, the breadth of stakeholders involved, and their depth of involvement in various stages of the evaluation. Evaluations also differ in the political unanimity or frictions encountered among stakeholders, the feasibility of implementing the study to obtain the degree and nature of participation desired, and the ultimate outcomes in terms of use or change in stakeholders. But each approach provides interesting ideas and options for the practicing evaluator to consider in the context of the individual evaluations that he or she undertakes. Our brief summary of each approach is intended to motivate you to read more about these approaches and consider the elements that may best fit the context of the program you are evaluating. In any case, participatory evaluation

is here to stay. Evaluators now realize that they differ from applied researchers and that they must learn about, and from, the stakeholders concerned with the project they are evaluating to implement and produce a more valid and useful evaluation.

How Participant-Oriented Evaluation Approaches Have Been Used

Part Two of this book is concerned with evaluation theories or approaches. In the preceding section, we have been concerned with describing and contrasting current participatory approaches. But we are also interested in how today's participatory approaches are used in practice. Several authors have addressed this issue. Their results indicate that many practicing evaluators involve stakeholders in their evaluations and consider at least some of their evaluations to be participatory. However, few use stakeholder participation as extensively as some of the theoretical approaches suggest.

Research on Involvement of Stakeholders

Christie's work (2003) illustrates the centrality of stakeholder involvement to evaluation theories and to those who work extensively in evaluation. Although her focus was not solely on participatory evaluation, her results are relevant here. Christie first selected eight prominent evaluation theorists and, using surveys to examine their views about evaluation, found that all eight evaluators, in spite of major differences in their theories or approaches to evaluation, involve stakeholders in their evaluations. Although the eight differed in their scores, all fell in the positive end of the dimension.[10] However, when Christie surveyed practitioners, a group of people who conducted evaluations for Healthy Start programs in California, she found that they were not as concerned with stakeholder participation as were the theorists. A majority (63%) of external evaluators made use of stakeholders, but internal evaluators were less likely to do so (43%). The internal evaluators, however, were generally not full-time evaluators and had other responsibilities in the organization.

Two studies, more than ten years apart, collected more extensive data on practicing evaluators' use of participatory or collaborative approaches. Cousins, Donohue, and Bloom (1996), in a survey of 564 evaluators from the United States and Canada,

[10]As a whole, the group represents a broad spectrum of evaluation approaches. Four of the eight evaluation theorists would be considered participatory in their approaches: House, Cousins, Patton, and Fetterman. The remaining four do not emphasize participation of stakeholders in their approaches (Stufflebeam, Boruch, Eisner, and Chen). Therefore, the fact that all scored in the positive end on the dimension of stakeholder participation is informative and suggests the extent to which participation has become embedded in evaluation theory and practice, at least among evaluation leaders, since the early years in the field.

found that 62% reported using a collaborative approach in a recent evaluation; however, their descriptions of their evaluation practices were most like the traditional stakeholder-based evaluative approach. In particular, the evaluator remained in control of the evaluation study and the most frequent use of stakeholders was for defining the scope of the evaluation (72%) and interpreting results (51%). Few involved stakeholders in technical tasks; these remained the realm of the evaluator.

In her research on participatory evaluation in international development programs, the settings in which many theorists advocate transformative participatory evaluation, Cullen concluded that "participatory evaluation approaches are not as widespread in development evaluation as the literature would suggest" (2009, p. 79). After collecting data from evaluation reports, surveys with 166 evaluators in 55 countries, and interviews with 15 evaluators, Cullen found that 72% of those surveyed reported having used a participatory approach in a recent evaluation. However, their definitions of participatory evaluation and their actions varied greatly—from evaluations that simply gave stakeholders a voice or used them as a data source to those that emphasized empowering stakeholders or capacity building. Like the evaluators in the study by Cousins et al. (1996), those surveyed and interviewed tended to maintain control of the evaluation themselves and were most likely to involve stakeholders in tasks such as collecting data and disseminating findings, but not in evaluation design or data analysis. Cullen found that a large proportion of those surveyed reported involving program staff (82%) and program recipients (77%); however, some of that involved participating in data collection as a respondent. In comments on the survey and in interviews, these evaluators expressed concern at the extent to which program or organizational donors dominated the participatory process. In these international development evaluations, at least, transformative evaluation approaches to change the power structure and empower oppressed recipients were rarely the focus. Other results suggested that these international development evaluators saw participatory approaches as most useful and most feasible when there was little conflict among stakeholder groups and the focus of the evaluation was formative.

These studies indicate that many practicing evaluators, from 62% to 72% in these three studies, involve stakeholders, but typically in more traditional ways. Although participative approaches, along with theory-based evaluation, dominate much of the current theoretical writing in evaluation, it appears that the world of practice differs from that of theory. Christie found that internal evaluators, who often have other program responsibilities and may not identify strongly with or be trained extensively in evaluation, often do not consider stakeholder participation, but the practices of external evaluators in her study were similar to those found in the other research. Cousins et al. and Cullen survey evaluation practitioners who are members of professional evaluation associations and, therefore, are more likely to identify themselves as evaluators and be well trained in the field. These evaluators often involve stakeholders and claim to use participatory approaches, but their definitions of participatory evaluation differed dramatically and their practices tended to mirror stakeholder-based participatory approaches rather than some of the newer approaches.

Perhaps more time is needed for the contemporary participatory approaches to be used frequently in practice. More likely, time will allow some of their concepts to be adapted when they are appropriate to the context of the organization and program and to the purposes of the evaluation. Many approaches, as with theories, are proposed by the authors, not with the intent to be practiced literally in the way the author describes, but to inspire evaluators and other users of evaluation to consider the concepts and their utility in building evaluation capacity or facilitating use.

Use of Approaches by Developers

Meanwhile, many authors of participatory approaches provide examples of the use of their approaches by themselves and others in their books and articles. For example, Weaver and Cousins (2004) provide an extended discussion of two cases in which they applied P-PE. They describe the cases and rate each on the dimensions Cousins has used in the past for differentiating participatory approaches. In addition, they add two new dimensions for considering participatory approaches: power relations among participating stakeholders (neutral or conflicting) and manageability of evaluation implementation (manageable or unmanageable). These address the common concern that participatory evaluation can be much more time-consuming and costly than a typical evaluation, particularly when involving many stakeholders or stakeholders with strong differences. Patton, in his writing on developmental evaluation, gives insight into the way in which he works with organizations using this approach (1997c, 2005b, 2010). Also, as noted earlier, Fetterman and Wandersman's latest book (2005) includes chapters by several authors who report on their use of empowerment evaluation in different settings. An interview with Jean King illustrates a participative approach for capacity building in which she, as the evaluator, acts as a facilitator while stakeholders make decisions. (See Fitzpatrick and King, [2009].) Finally, four chapters in a *New Directions for Evaluation* issue on deliberative democratic evaluation present cases by eminent evaluators on their attempts to promote inclusion (Greene, 2000); use dialogue (Ryan & Johnson, 2000; Torres, Padilla Stone, Butkus, Hook, Casey, & Arens, 2000); and incorporate inclusion, dialogue, and deliberation (MacNeil, 2000) in evaluations. Green frankly discusses the problems she encountered in attempting to achieve inclusion in discussions and evaluations of a controversial high school science program. Torres et al., discuss how they and the coordinators of the program being evaluated used dialogue and deliberation to achieve greater understanding and learning about evaluation. Ryan and Johnson's dialogues occur with faculty, students, and deans, often in separate groups, to consider course evaluations and learning. MacNeil describes a fascinating case in a locked psychiatric institution where she attempted to carry out a participative evaluation with dialogue among many groups. She provides useful and frank advice on her successes and failures in establishing and organizing forums for dialogue.

Strengths and Limitations of Participant-Oriented Evaluation Approaches

As the many contemporary participative models described in this chapter demonstrate, participative approaches are alive and well in program evaluation theory, if less so in practice. The approaches vary so dramatically that it is difficult to summarize their strengths and weaknesses as a group. We have mentioned some limitations in reviewing each model. Here, we will summarize some of the trends.

Strengths of Participatory Approaches

Participatory approaches have certainly made evaluators aware of the potential value of including stakeholders in the process of conducting a study, not simply as data sources or to assist in distributing surveys or for setting up focus groups, but for improving the validity of the study and its use. How? Involving stakeholders in thinking about the meaning of various constructs or phenomena to be measured, conceptualizing them, and considering how to measure them or collect information on them can result in more valid data. Why? Because as Stake realized long ago, different stakeholders have knowledge and perspectives that evaluators do not have. They know the program and its context. Teachers, counselors, parents, and students can all be invaluable in helping evaluators understand what a particular school means by students "at risk." Evaluators' dialogue with them increases the evaluators' understanding of the concept and, ultimately, the selection or development of better, more valid measures for the evaluation. Similarly, involving the stakeholder in the study in this way helps them gain trust in the evaluation, begin to understand it, and consider how they might use it.

Evaluation Is Always a Partnership. Evaluators have the skills to analyze the merit or worth of a program and to make recommendations for improvement or action. But, stakeholders have knowledge that we do not. Policymakers know what other programs they are considering, the budgetary and time constraints for their decisions, and the factors, political and otherwise, that may influence their decisions. Program managers and staff know the details of the program and some of the struggles and successes of the students or clients. They know what they have tried before and why it didn't work. They know the logic model for the current program, why they adopted it and what elements of it they think may work. They have expertise in working with these types of clients or types of problems. Finally, the clients themselves, of course, know the details of their own lives, the problems they struggle with and the solutions they find, and how they respond to and think about the program or policy. Many evaluations make use of advisory groups of stakeholders to pursue this partnership—to exchange information on the

program, the clients, and the evaluation. Participatory evaluation, especially the practical approaches, is simply extending or modifying that effort to learn more from stakeholders and to have them learn more about the evaluation.

Practical participatory methods are intended to increase use as the stakeholders begin to understand the evaluation, make decisions about it, and ultimately be excited about the results. (See interview with Fitzpatrick and Bledsoe [2007] for an example of an evaluation in which program managers are moved to unanticipated use through the evaluator's work in involving them in the evaluation, thereby gaining their trust and interest.) Some authors have reported research findings showing greater use when stakeholders are more involved (Cullen, 2009; Greene, 1988; Leviton & Hughes, 1981; Patton, 1986; Weiss & Bucuvalas, 1980b). But others (Henry & Mark, 2003) have argued that the evidence is more limited. Certainly, more research is needed in this area to examine the types of participation and their possible linkages to types of use. A word on that: Deep, participatory involvement of a few managers and staff, as found in practical participatory evaluation and empowerment evaluation, is probably most successful, as Cousins and Earle explicitly acknowledge, in formative evaluations—those concerned with program improvement. Program managers and staff are rarely those who make summative decisions about program continuation, though they may make decisions about expansion, for example, whether to expand a pilot to other classrooms, other students, or other similar schools. Nevertheless, summative decisions are often made by policymakers who are higher in the organization, central office managers, or school board members who make budgetary decisions. These individuals are not likely to have the time to become closely involved in the evaluation of a program, though they may send a representative if the program is extremely costly or controversial, and that representative's participation may influence use. Stakeholder participation is most likely to lead to greater use when the intended primary users are managers and/or program staff who will use the information in a formative way. In such cases, the deep involvement of these stakeholders can certainly carry the added advantages of building program capacity and organizational learning, moving those who work in the organization toward more data-based decisions.

We will talk further about organizational learning and capacity building in Chapter 9. These concepts are certainly intended outcomes of some participatory approaches, but are not participatory approaches themselves, which is our subject here.

Limitations of Participatory Approaches

What are the drawbacks of participatory approaches? Well, there are many, and an evaluator will have to balance the advantages and disadvantages of undertaking a participatory approach in each new setting, considering the nature of the program, the organization, and the potential stakeholders to involve. The two broad categories of drawbacks are (a) the feasibility, or manageability, of implementing a successful participative study; and (b) the credibility of the results to those who do not participate in the process. Regarding feasibility, as noted, Weaver and Cousins

(2004) added manageability as a dimension to consider in judging different collaborative approaches because achieving stakeholder participation is rarely an easy task. Cullen's survey of 166 evaluators who used participatory approaches to evaluation in international development settings in 55 different countries found that, although there were great benefits (with 93% reporting that the participation resulted in more useful information), a majority of the evaluators reported drawbacks in the increased time (69%) and costs (58%) of undertaking a participatory evaluation. The deliberative democratic approach was developed as an ideal, but some have given it a try to explore its feasibility. They encountered particular difficulties in obtaining participation from the most disadvantaged stakeholders. (See Greene [2000] for her discussion of her failed attempts to gain the participation of disenfranchised groups, those with less power, in a high-profile and politically charged evaluation of a new high school science program.)

Other concerns regarding feasibility have to do with evaluators' skills. Evaluators are trained in research methods. Although some of that training concerns working with stakeholders or teams to identify evaluation questions or disseminate results, few evaluators have extensive training or skills in facilitation, particularly with groups that may have political differences. (See, for example, some of the personal problems Fetterman encountered in evaluating a program in his own school. Faculty colleagues and his dean presented major, unanticipated problems and criticism. [Fitzpatrick & Fetterman, 2000].) Laura Leviton, in her presidential address at AEA, spoke to the need for evaluators, more inclined to analytical skills and working alone, to develop "people skills" to work with others, including stakeholders new to evaluation. As she indicates, there is much research on communication skills and conflict resolution that evaluators could use, but our professional literature tends to focus on the analytical. Not only does stakeholder participation take time and resources, but acting as a facilitator, coordinator, and consultant to stakeholders on an evaluation requires good listening and facilitation skills.

The other main area of concern with participatory evaluation is the credibility of the results to those who do not participate in the study. When stakeholders, particularly those delivering or managing the program, are heavily involved in the evaluation, other audiences wonder if these stakeholders are able to be objective about their own programs. There is a potential for bias. Bias can occur even within the evaluator as he or she begins to know the stakeholders well and is co-opted by them. Scriven (1973) developed goal-free evaluation at least partly to minimize contact between the evaluator and program people and thereby reduce the bias introduced by personal relationships. (See Fitzpatrick and Bickman [2002] where Bickman also expresses his desire to avoid contact with program people to improve the credibility and legitimacy of the study. His client was the funding agency for the evaluation.) Others are concerned with whether stakeholders can refrain from bias when they are heavily involved in the evaluation of their own program. It is difficult to judge one's own work objectively. That is why funders and even organizations themselves tend to hire external evaluators for important, controversial, or costly programs. Their view is that the newness, the lack of ownership, that external evaluators bring will enable them

to conduct an evaluation that contains less bias or at least less bias regarding the program itself. (The evaluators, as noted, bring their own experiences and views that introduce other sources of bias.)

A final important concern is the competence of stakeholders to perform the tasks that some approaches call for. Peter Dahler-Larson (2006) speaks to this concern in regard to the "popularization" of evaluation, which has led organizations to want to know how to do evaluations themselves. In many ways, this is good. Don't we want organizations to improve their performance, and don't we think evaluative ways of thinking will help them do so? However, some participatory approaches such as empowerment evaluation (in theory) and transformative participatory approaches give the stakeholders ownership of the evaluation, with the evaluator acting as a technical consultant. Dahler-Larsen notes that it takes years of training and practice in many different settings for evaluators to become good at what they do. He bemoans the "expectation that lay persons will become good evaluators based on a short introduction and rudimentary training in methodology" (Dahler-Larsen, 2006, p. 145). Evaluators themselves often call in others, either on their team or outside evaluators, who have expertise in other areas, but these approaches appear to assume that the stakeholders will do it all, with a little consultation from the evaluator. "Other strategies for bridging the competence gap are to make the best of the situation, 'get a process going,' do some 'capacity building,' and hope, sometimes desperately without much justification, that the evaluation will be even 'better next time'" (Dahler-Larsen, 2006, p. 145).

More than any other approach to program evaluation that we have examined, participant-oriented approaches add a political element inasmuch as they foster and facilitate the activism of recipients of program services. But, as we acknowledged in Chapter 3, politics is part of evaluation and the good evaluator embraces that involvement, recognizing that evaluation provides one piece of information to help improve government and society. Today many citizens seek an active role in governance. Parents have become active participants in school policy making. Citizens are involved in crime-prevention programs and in developing programs to improve their environment and their communities. Programs intended to serve AIDS patients, the homeless, the chronically mentally ill, and tenants in public housing programs are also carved out of a politicized climate. Programs such as these often receive much national and local media attention. In today's participatory culture, the participatory approaches we have reviewed provide a means for empowering and educating stakeholders, particularly those with less access to power, to make informed choices and become involved in decision making.

Major Concepts and Theories

1. In participant-oriented evaluation approaches, evaluators involve the stakeholders of the program, policymakers, program sponsors, managers, staff, program recipients, or others with a stake in the program in the planning and implementation of the evaluation.

2. Robert Stake's Countenance Framework and Responsive Evaluation and Guba and Lincoln's Fourth-Generation Evaluation and naturalistic approaches emerged in the 1970s and 1980s in response to the dominance of quantitative, preordinate, social science methods focused on causality that were the primary evaluation methods at that time. Stake, and Guba and Lincoln wanted methods and approaches that were more responsive to stakeholders' needs, which may not necessarily focus on establishing causality; more flexible and adaptive to change during the evaluation to reflect changes in the program and in the evaluator's knowledge; and more successful at providing an in-depth description and understanding of the program to the audiences for the evaluation. They acknowledged the multiple realities of a program and the need for the evaluator to portray those multiple realities as seen by different stakeholder groups.

3. Today's approaches to participative evaluation may be divided into two streams that have different primary purposes. Practical participative approaches involve stakeholders in an attempt to increase use of the evaluation results based on the view that involvement gives stakeholders, the potential users, a sense of ownership and understanding that increases the likelihood of use. Transformative participatory approaches involve stakeholders for political reasons: to empower those stakeholders by providing them with tools in evaluation and self-determination or insights into power arrangements in their community or society, and tools to change those power arrangements.

4. Each participatory approach may be categorized along three primary dimensions: (a) whether the evaluator or stakeholders have primary control over the decisions regarding the evaluation, (b) stakeholder selection or the breadth of stakeholders involved in the study, and (c) depth of stakeholder involvement. Two more recent dimensions include (d) power relations among stakeholders (conflicting or neutral) and (e) manageability of the study, that is, ease or difficulty in implementing stakeholder participation in the desired manner.

5. Well-known practical participative approaches include stakeholder-based evaluation (Mark & Shotland, 1985), practical participatory evaluation (P-PE) (Cousins & Earl, 1992), and developmental evaluation (Patton, 1994, 2010). The stakeholder-based approach retains evaluator control while seeking input from a broad variety of stakeholders on key issues of evaluation purposes, focus, and final interpretation and dissemination. Practical participatory evaluation and developmental evaluation use shared decision making or partnerships between the evaluator and some key stakeholders, often program managers or staff, who are involved in depth in the evaluation or developmental activities.

6. Well-known transformative approaches include empowerment evaluation and deliberative democratic evaluation. Although these approaches differ dramatically in their participative dimensions, they are transformative in that they address political purposes. Empowerment evaluation (Fetterman, 1994, 1996, 2001a) is intended to empower the stakeholders involved. It gives primary decision-making power to them, and involves these stakeholders, often program managers and staff, deeply in the evaluation to provide them with the skills to conduct evaluations in the future and thus become self-determining. Deliberative democratic evaluation (House and Howe, 1999) is founded on the democratic values of social equity and equality and is concerned with addressing the typical power imbalances among stakeholders to permit each stakeholder group, especially those least likely to give input, to participate. Decisions regarding the evaluation remain in the control of the evaluator, but all legitimate stakeholder groups are involved in the processes of dialogue and deliberation to help the evaluator learn the

various, real needs and views of stakeholders. Empowerment evaluation has been practiced and described in many different situations. Deliberative democratic evaluation is intended more as an ideal model to guide evaluators' practice.

7. Research on the use of participative approaches suggests that, in practice, evaluators involve stakeholders (Cousins, Donohue, & Bloom, 1996; Cullen, 2009; Fetterman & Wandersman, 2005; Miller & Campbell, 2006), but definitions of what constitutes participative evaluation differ dramatically and generally do not follow the tenets of a particular model. Many theorists encourage adaptation of the model to the local context.

8. Strengths of the participative approaches include their potential to increase understanding and use of evaluation by stakeholders and to increase evaluators' understanding of programs and organizations and, in so doing, to provide more valid and useful information. Stakeholder participation can also lead to organizational learning. As those in the organization learn more about evaluation, they may use it or its modes of thinking and using data to address problems in the future. Weaknesses include concerns about bias and, therefore, less acceptance of the study by external audiences; greater time and cost to involve stakeholders; potentially weaker results if those conducting the study lack necessary skills; and the implication that skills and expertise in evaluation can be readily and quickly gained by any stakeholder.

Discussion Questions

1. When would a participant-oriented evaluation approach be particularly appropriate? Provide a few examples from your own experience.

2. What kinds of risks does an evaluation entail by involving many different stakeholder groups?

3. Compare and contrast participant-oriented evaluation with objectives-oriented evaluation. What are the major differences between these two approaches? What are the similarities?

4. What do we mean by multiple perspectives of a program? Why would different groups have different perspectives? Shouldn't the evaluation just describe the one real program?

5. Is it useful for a manager to know about different perspectives of a program? Why? What can a manager (or other stakeholders) do with such information?

6. Which participative approach is most appealing to you? Why? What do you see as its strengths and weaknesses?

7. Which participative approach might be most useful to evaluate a particular program in your organization? Why would this approach be useful? What would be the risks in using the approach?

8. Who are the common stakeholders for evaluations in your work? Discuss their interest and capacity in being involved in an evaluation. (See Fitzpatrick and King [2009] for an example of an evaluation in which many school district personnel and parents actively participated.)

Application Exercises

1. Consider a program that you are involved with that would benefit from a participant-oriented evaluation. What would be your purpose in using this approach in your evaluation? Which approach would you select and why? Consider how you would select stakeholders, the ways in which you would share or not share decisions with them, and the manner in which you would involve them. Discuss the areas in which you would follow the principles or steps of the approach and areas in which you would deviate.

2. As newly appointed director of student activities for the John F. Kennedy High School, you decide to conduct an evaluation of the student activities program in the school. The most current information about the program is found in the faculty handbook, published at the opening of each school year. This description reads as follows: "The John F. Kennedy High School offers a wide range of activities for its 2,500 students. Among the various activities are clubs, intramural and varsity sports, band, choir, orchestra, and various service programs such as Red Cross. Clubs are organized by students and assigned a faculty advisor by the dean of students. Meetings are scheduled on Monday to Thursday evenings and held in the cafeteria, auditorium, or gymnasium of the school. Varsity sports activities are directed by members of the physical education faculty. Intramural sports are organized by home rooms and directed by a faculty member appointed by the dean of students. Band, choir, and orchestra are under the direction of members of the music department. Service programs are organized by students who must also find a faculty member who is willing to advise them." This description does not provide you with sufficient insight into the program, so you decide to conduct an evaluation of the current program before undertaking any modifications or restructuring of the program. As a participant-oriented evaluator, how would you proceed to plan and conduct the evaluation?

3. Describe the similarities and differences among responsive, practical participatory, and empowerment evaluation approaches.

4. Read the two participatory examples discussed in Weaver and Cousins (2004) and two of the chapters illustrating the implementation of empowerment evaluation in Fetterman and Wandersman (2005) and contrast the actions of the evaluator and stakeholder in these evaluations. How are they alike? How do they differ? How are the actions and roles like or unlike the approach they are intended to represent?

5. **a.** In your own work, identify two evaluation questions that would be most appropriately answered with a participant-oriented approach. Identify two questions or issues that might be inappropriate to address with a participant-oriented approach. Why do you think the first questions lend themselves to a participant-oriented approach and the second ones do not?
   b. Choose one of the questions you identified as appropriate to address with a participant-oriented approach. What stakeholder groups would you involve in the evaluation? How would you involve them? At what stages? Would you take a case study approach? How would your data collection differ from that used in a more objective approach?

6. Use a participatory approach to evaluate the following program: Arvada Hospital has begun an aggressive program to screen men who are over fifty years of age for prostate cancer. Their first efforts involved a saturation advertising program using billboards and radio and newspaper advertisements with well-known local men advocating screening and describing their experience. The hospital has hired you to use a participant-oriented approach to evaluate this effort. What do you do? What stakeholders do you involve? How?

Case Studies

We recommend two interviews that make use of participatory elements for this chapter: In *Evaluation in Action*, Chapter 9 (Jean King) and 11 (Ross Conner).

In Chapter 9, Jean King discusses how she uses a participatory approach to evaluation capacity building (ECB) in which her role is primarily as a facilitator while other stakeholders, primarily teachers and parents, make the critical decisions on the evaluation of a highly political special education program in a school district. This chapter is only available in the book.

In Chapter 11, Ross Conner describes how he works with the funder and citizens of different communities in Colorado in many roles, including evaluator, to help them build the program and evaluate it. This chapter was expanded from the original interview, but portions of the interview can be found here: Christie, C., & Conner, R. F. (2005). A conversation with Ross Conner: The Colorado Trust Community-Based Collaborative Evaluation. *American Journal of Evaluation*, *26*, 369–377.

Suggested Readings

Cousins, J. B., & Earl, L. M. (Eds.). (1995). *Participatory evaluation in education*: *Studies in evaluation use and organizational learning.* Bristol, PA: Falmer Press.

Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. In E. Whitmore (Ed.), *Understanding and practicing participatory evaluation.* New Directions for Evaluation, No. 80, 5–23. San Francisco: Jossey-Bass.

Fetterman, D. M. (1994). Empowerment evaluation. *Evaluation Practice, 15*, 1–15.

Fetterman, D. M., & Wandersman, A. (Eds.). (2005). *Empowerment evaluation principles in practice.* Thousand Oaks, CA: Sage.

Fitzpatrick, J. L., & King, J. A. (2009). Evaluation of the special education program at the Anoka-Hennepin School District: An interview with Jean A. King. In J. Fitzpatrick, C. Christie, & M. M. Mark (Eds.), *Evaluation in action* (pp. 183–210). Thousand Oaks, CA: Sage.

Greene, J. C., & Abma, T. A. (Eds.). (2001). *Responsive evaluation*. New Directions for Evaluation, No. 92. San Francisco, CA: Jossey-Bass.

Patton, M. Q. (1994). Developmental evaluation. *Evaluation Practice*, 15, 311–320.

Patton, M. Q. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation and use.* New York: Guilford.

Ryan, K. E., & DeStefano, L. (Eds.). (2000). *Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation.* New Directions for Evaluation, No. 85. San Francisco, CA: Jossey-Bass.

9

Other Current Considerations: Cultural Competence and Capacity Building

Orienting Questions

1. What do we mean by cultural competence in evaluation? Why is it important to have cultural competence in the context of the program you are evaluating?
2. What effects can evaluation have on an organization?
3. What do we mean by mainstreaming evaluation or evaluation capacity building? Is it desirable?
4. How do we go about building evaluation capacity?

Before we move to a discussion and comparison of the different approaches we reviewed in the last few chapters, we want to discuss two factors that influence our evaluation practice today, but transcend particular evaluation approaches. These are (a) our growing awareness of the need for evaluators to consider and build cultural competence for the program they are evaluating, and (b) our role in influencing not just programs, but organizations by evaluation capacity building or mainstreaming evaluation. Regardless of which approach or parts of approaches we are using, our evaluations take place in both an organizational context and a broader cultural context that influences or should influence our choices. So before we move to Parts Three and Four of this book that focus on how to conduct evaluations, we want to discuss cultural competence and capacity building. If we do not consider these issues from the beginning of our evaluation, the validity, power, and use of the evaluation will be diminished. In fact, cultural competence and skills for building evaluation capacity are part of the how-to of evaluation.

The Role of Culture and Context in Evaluation Practice and Developing Cultural Competence

Growing Attention to the Need for Cultural Competence

In Karen Kirkhart's 1994 Presidential Address to the American Evaluation Association, she observed that our "multicultural influences shape and are shaped by our work" and argued that "multicultural validity," a term she introduced in her address, should "be conceptualized as a central dimension of validity, treated with the same respect, routinization, and scrutiny as other dimensions; that is, it should be a visible focus of concern in evaluation, theory, methodology, practice, and metaevaluation" (Kirkhart, 1995, p. 1). Since then—with a few ups and downs—evaluators have begun to consider how their work is influenced by culture and the need for evaluators to understand the cultures in which they work, and in so doing, to improve the validity and utility of their evaluations.

Her call to attend to multicultural influences promoted much discussion, organizing, writing, and action, though we now recognize, as Kirkhart indicated, that multicultural understanding is "a journey." But some important actions have been taken. Revisions of the *Guiding Principles* of the American Evaluation Association in 2004 addressed the importance of cultural competence under the existing category of expected competencies for evaluators:

> To ensure recognition, accurate interpretation and respect for diversity, evaluators should ensure that the members of the evaluation team collectively demonstrate cultural competence. Cultural competence would be reflected in evaluators seeking awareness of their own culturally-based assumptions, their understanding of the worldviews of culturally-different participants and stakeholders in the evaluation, and the use of appropriate evaluation strategies and skills in working with culturally different groups. Diversity may be in terms of race, ethnicity, gender, religion, socioeconomics, or other factors pertinent to the evaluation context. (Accessed January 12, 2010, from www.eval.org/Publications/GuidingPrinciples.asp)

More recently, the Joint Committee on Standards for Educational Evaluation approved several new standards that include an emphasis on cultural competence:

- **U1** *Evaluator Credibility* (emphasizing the need for the evaluator to have credibility in the context of the program and the evaluation)
- **U4** *Explicit Values* (clarifying and specifying cultural values underpinning purposes, processes, and judgments in the evaluation)
- **F3** *Cultural Viability* (recognizing, monitoring, and balancing cultural and political interests and needs)
- **P2** *Formal Agreements* which take into account cultural contexts
- **A1** *Justified Conclusions and Decisions* (justified in the culture and context where they have consequences), and others. (See Appendix A for the full text of each of these standards.)

The emphasis on cultural competence that suffuses this revision of the *Standards* attests to evaluators' recognition of the importance of context and culture on many phases of the evaluation.

Other evidence of the recognition of cultural competence and its impact abounds. The National Science Foundation has developed a handbook that describes the essentials of cultural competence in evaluation (Frierson, Hood, & Hughes, 2002; Frechtling, 2002). In addition, there have been ongoing discussions and several publications on cultural competence. (See Botcheva, Shih, and Huffman [2009] and Thompson-Robinson, Hopson, and SenGupta [2004].)

We will introduce the reader to some of the issues of concern in this area by discussing what cultural competence is, why it is particularly important to evaluation, and what benefits will emerge from conducting an evaluation with cultural competence. SenGupta, Hopson, and Thompson-Robinson (2004) define cultural competence in evaluation as follows:

> A systematic, responsive inquiry that is actively cognizant, understanding, and appreciative of the cultural context in which the evaluation takes place; that frames and articulates the epistemology of the evaluative endeavor; that employs culturally and contextually appropriate methodology; and that uses stakeholder-generated, interpretive means to arrive at the results and further use of the results. (2004, p. 13)

Both AEA's Guiding Principle on competence and SenGupta et al.'s definition emphasize the evaluator's obligation to become knowledgeable about and to understand and appreciate the culture of the program to be evaluated and the context in which it operates. But we often are not even cognizant of our own cultural norms; they may be all we know and are certainly what we know best. (Does the fish know it's swimming in water?) As the AEA Guiding Principle indicates, we must first become more cognizant of our own cultural norms and values, and how they affect our perspectives. Of course, they differ for each of us. But with this self-awareness, we can begin to learn the cultural norms, values, and accepted behaviors for the culture of the program we are evaluating.

Why Is Cultural Competence Important?

As a group, evaluators, at least in the United States, are not a very diverse lot other than in gender. An environmental scan of members of the American Evaluation Association conducted in 2007 showed that, of the 2,637 responding members, 67% were women and 73% were white.[1] (Accessed January 11, 2010, from http://www.eval.org/Scan/aea08.scan.report.pdf.) The United States has become a multi cultural, pluralistic country, and evaluators do not represent that pluralism. Yet, we often evaluate school, government, or nonprofit programs that serve a

[1]No other single racial or ethnic group dominated, nor were they in proportion to their prevalence in the U.S. population: 8% were members from other countries, 7% were Black or African American, 5% were Asian, another 5% were Hispanic or Latino, and 2% were American Indian or Alaskan.

significant number of people whose culture—experiences, views, perspectives, and behaviors—differ from our own. In some cases, the stakeholders served by the program are also immigrants, new to our country and our culture. Finally, many of us conduct evaluations in other countries whose cultures we do not know. So, many, many evaluators are studying programs, collecting data from students or clients in those programs, interpreting that data, and making recommendations for program change even though we are not familiar with the culture and context in which many of the clients live. This lack of familiarity affects our ability to appropriately direct and guide the evaluation, to conceptualize variables and ways to collect data on them, to analyze and interpret results, and to disseminate them to these stakeholder clients in ways they can understand and use. As Stafford Hood has written, "It is difficult, if not impossible, for evaluators to see, hear, and understand cultural nuances that differ from their lived experience" (2005, p. 97). Recall that Robert Stake (1975b) was one of the first to write about the need for evaluators to be responsive to the context and to portray an accurate, full descriptions that helped others to understand a program. Hood is helping us realize that we cannot do that without a better understanding and appreciation of the cultures represented in the programs and contexts we study.

Because evaluators are typically not of the same culture as program participants, gaining cultural competence is important to the evaluation for a variety of reasons:

- To help the evaluator identify the issues of concern to stakeholders and incorporate them into the purposes of the evaluation
- To consider, from multiple perspectives, what constitutes credible evidence of program success
- To collect valid data and information, sensitive to the norms, perspectives, and beliefs of the groups from whom data are collected
- To analyze and interpret data to achieve valid results
- To increase the legitimacy of the evaluation to all stakeholders, including those new to or fearful or skeptical of evaluation and its methods
- To disseminate the results in ways that are understandable and responsive to each stakeholder group's norms and values
- To increase the usefulness and use of the results
- To recognize problems with inequality or power and the opportunities to improve social equity and democratic values through the evaluation and the way in which it is conducted

As the reader can see, cultural competence is not just a nice extra to an evaluation. It is critical to developing and completing effective evaluations, ones that make a difference in programs and in society. Another fruitful role for the culturally competent evaluator is in program planning. As we have noted, many approaches (CIPP, developmental evaluation, transformative evaluation) encourage evaluators not to wait until the end of the program to assess performance, but to collect evaluative data during the planning stage so as to lead to programs that are more effective, at least partly because they are responsive to the culture of the community. (See Mertens [1999].)

How can we achieve cultural competence? As noted, it is a journey. Improving one's cultural competence for an evaluation is not nearly as easy as learning a new statistical test, because there are no clear rules or steps that fit every setting and every evaluator. However, here are some useful suggestions that emerge from the literature:

- Be aware of how your own culture—everything from where you grew up; the racial, ethnic, and religious heritage of your family; and your customs and beliefs to your university training and your practice as an evaluator—affects your beliefs, values, and behaviors.
- "Practice constant self-examination of values, assumptions, and cultural contexts" (SenGupta et al., 2004, p. 13).
- Particularly when coming to the program or meeting with stakeholders for the first time, use the opportunity to learn through quiet observation, respectful interactions, and reflection.
- As the culture becomes increasingly different from your own, make inclusion an essential priority (Madison, 2007). Seek participation and input from stakeholders who represent the different culture. Develop relationships with them to learn more.
- Have at least one member on your evaluation team who represents that culture in some way. That member should serve important roles in communication and interpretation.
- As a member of the evaluation professional community, work to recruit evaluators who represent different cultures, and racial and ethnic groups (Hood, 2001, 2005).

See Wadsworth (2001) for a useful example of a responsive evaluation and her description of dialogues with culturally different groups and learning more about local communities and their views.

Evaluation's Roles in Organizations: Evaluation Capacity Building and Mainstreaming Evaluation

Michael Patton was one of the first to recognize the impact of evaluation on organizations. He developed the concept of "process use" to illustrate that impact. Process use refers to "individual changes in thinking and behavior, and program or organizational changes in procedures and culture, that occur among those involved in evaluation as a result of the learning that occurs during the evaluation process" (Patton, 1997c, p. 90). Evaluators began to recognize that the impact of their evaluation work was not only in the results themselves, but also in the learning that occurred about evaluation and evaluative ways of thinking among the employees in the organization. This learning then influenced the organization itself. Participative approaches to evaluation increased that learning, at the very least by involving more employees. Sometimes the organizational changes that occurred were, in fact, a primary goal of the evaluation (for example, in some transformative approaches).

A related factor that prompted evaluators to think more about evaluation's role in organizations was the increased demand or popularization of evaluation in organizations. (See Dahler-Larsen [2006].) Requirements for evaluation, often for accountability purposes, have increased in the last decade. With those requirements have come increased responsibilities for agency employees to conduct evaluations or collect and report evaluative data. (See Christie and Barela [2008] Datta [2006] and Hendricks et al. [2008] for data and discussions of evaluations conducted by nonevaluation personnel in organizations.)

Both the recognition that evaluation has an impact beyond programs, on organizations themselves and on their ways of making decisions, and the increasing involvement of program managers and staff in evaluations have prompted evaluators to begin thinking and talking more about this impact on organizations. Two related discussions that have emerged are evaluators' attempt to "mainstream" evaluation in organizations (Sanders, 2002; Barnette & Sanders, 2003) and to build evaluative capacity in organizations.

Mainstreaming Evaluation

As we discussed in Chapter 2, mainstreaming evaluation is a concept developed by Sanders in his leadership of the American Evaluation Association in 2001. He describes it as follows:

> [It is] the process of making evaluation an integral part of an organization's everyday operations. Instead of being put aside in the margins of work, evaluation becomes a routine part of the organization's work ethic if it is mainstreamed. It is part of the culture and job responsibilities at all levels of the organization. (Sanders, 2002, p. 254)

Paul Duignan discusses the ways he and his colleagues in New Zealand, often a leader in evaluation, have worked to mainstream and to build evaluation capacity in organizations there. As he notes, in a small country like New Zealand, evaluators and other professionals are forced to become "multiskilled generalists" (2003, p. 8). He argues that "the most useful strategy in attempting to mainstream evaluation is probably to try and give it away" (2003, pp. 12–13). By giving evaluation away, he means having evaluation become part of the skills of the "multiskilled generalist." Evaluators there help others gain evaluation skills and help organizations build their organizational capacity for evaluation. Relevant to our discussion of approaches to evaluation, Duignan lists some characteristics of approaches that are most effective in mainstreaming evaluation or making it accessible to those in the organization for everyday decision making. A good approach, he argues, is one that does the following:

1. Demystifies evaluation.
2. Uses a set of evaluation terms that emphasizes that evaluation can take place right across a program's life cycle and is not limited to outcome evaluation.
3. Allows a role for both internal and external evaluators.

4. Has methods for hard-to-evaluate, real-world programs, not just for ideal, large-scale, expensive, external evaluation designs.
5. Does not privilege any one meta-approach to evaluation (for example, goal-free, empowerment) (Duignan, 2003, p. 15).

Similarly, Wandersman et al. (2003) describe a system they used to mainstream evaluation in statewide school readiness programs in South Carolina. They worked with programs across every county in the state, using some of the Getting To Outcomes (GTO) guides from empowerment evaluation and adapting them to the school settings. The results allowed the programs not only to meet accountability demands, but also to mainstream evaluation to use for internal decisions on program planning and improvement. In this way, Wandersman, Duignan, and others demonstrate ways that evaluators have begun to think about evaluation's impact on organizations and how to make evaluations part of the everyday decision-making process.

Evaluation Capacity Building

Others have written about organizations and evaluation with a focus on evaluation capacity building (ECB), which is defined as:

> A context-dependent, intentional action system of guided processes and practices for bringing about and sustaining a state of affairs in which quality program evaluation and its appropriate uses are ordinary and ongoing practices within and/or between one or more organizations/programs/sites. (Stockdill, Baizerman, & Compton, 2002, p. 8)

Stockdill and her co-authors note that ECB is not traditional evaluation and many evaluators will never be involved in it. But they argue that ECB can and should be a legitimate role for evaluators. Evaluation in organizations can have two forms:

(a) Traditional, project-based evaluation studies designed primarily to provide information on the program or policy being evaluated.
(b) ECB evaluation efforts which are an ongoing process within the organization, with the evaluator or staff person working to sustain an environment conducive to evaluation and its use within the organization.

ECB work, they note, is even more context-based than traditional evaluations because the ECB evaluator must consider issues regarding the history, structure, culture, and context of the organization; its decision-making styles, history, and so forth; as well as the context of the program to be evaluated.

Many organizations, among them the World Bank (Mackay, 2002), the American Cancer Society (Compton, Glover-Kudon, Smith, & Avery, 2002), and the U.S. Centers for Disease Control and Prevention (Milstein, Chapel, Wetterhall, & Cotton, 2002), have successfully based their evaluation efforts on organizational change and capacity building. (The references cited are examples of ECB at work in these

organizations. See also Jean King [2002] describing her work in ECB in a school district in Minnesota.) The point is to draw attention to these new roles—to build an organization's capacity, receptiveness, and use of evaluation—as appropriate activities that evaluators, often internal evaluators, evaluation managers, or other organizational staff, might undertake that are quite distinct from conducting evaluation studies. The Evaluation Capacity Development Group (ECDP) is an excellent source of useful information and examples of capacity building. See their web site and materials at www.ecdg.net.

Efforts and writing on capacity building and mainstreaming evaluation continue. Preskill and Boyle (2008) cite many catalysts for ECB, including Laura Leviton's American Evaluation Association 2000 theme of "Evaluation Capacity Building," Jim Sanders' 2001 theme of "Mainstreaming Evaluation," and the increasing use of stakeholders and participatory models in evaluation. In 2006, the American Evaluation Association created a new Topical Interest Group (TIG) on Organizational Learning and Evaluation Capacity Building, which has continued to increase interest in the topic. Furthermore, ECB efforts appear to be relatively common among evaluators. A scan of AEA members in 2007 found that 54% reported undertaking ECB efforts (American Evaluation Association, 2008).

Preskill and Boyle (2008) have added to our understanding of ECB through their development of a multidisciplinary model of ECB. (See Figure 9.1.) The circle on the left depicts ECB as it is initiated and implemented. The goal of ECB is reflected in the outer circle: increasing knowledge, skills, and attitudes about evaluation among employees in the organization. The pursuit of ECB in the organization is influenced by factors listed in the next circle: motivations, assumptions, and expectations about evaluation, ECB, and what it can accomplish. Finally, ten ECB strategies are listed in the inner most circle: internships, written materials, technology, meetings, appreciative inquiry, communities of practice, training, involvement in evaluation, technical assistance, and coaching. Each of these strategies is designed, implemented, and evaluated (see ring around the strategies with arrows connecting the stages) to build evaluation capacity. The factors shown in this circle then lead to transfer of learning (middle arrow) that connects ECB efforts to sustainable evaluation practices in the organization. In other words, the right circle might be viewed as an organization that has mainstreamed evaluation or has successfully built evaluation capacity to influence organizational thinking and practices. That influence, admittedly idealized, is illustrated in each of the circles within the larger sustainable evaluation practice circle. In other words, the organization with sustainable evaluation practices has continuous learning about evaluation, shared beliefs and commitment to evaluation, strategic plans for evaluation, and more. Both these circles, then, affect and are affected by the organization's learning capacity, leadership, culture, systems and structures, and communication practices. (See ovals between the two circles.) Ultimately, the results of these organizational ECB efforts are diffused to other organizations and individuals. (See arrow at bottom of model.) The figure, while complex and comprehensive, prompts us to consider each of these elements in implementing ECB within an organization. Preskill and Boyle, who see ECB as the future of evaluation, elaborate on each of these features in their

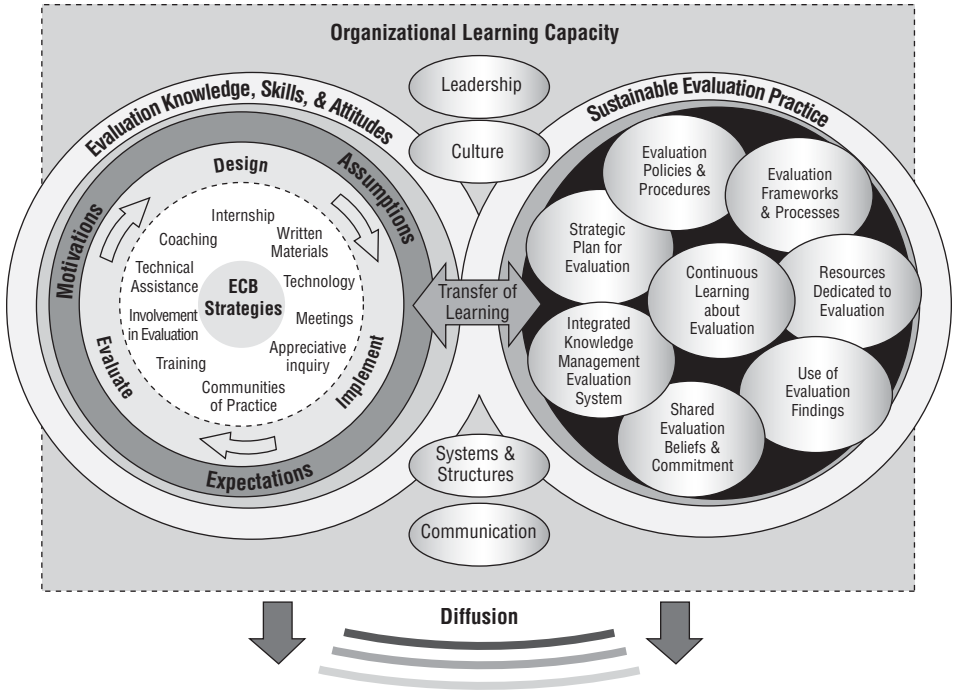


FIGURE 9.1 ***A Multidisciplinary Model of Evaluation Capacity Building (ECB)***

Source: From Preskill, H., & Boyle, S. (2008). A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation, 29*(4), 445. Republished with permission from Sage Publications.

article. They foresee evaluation moving from specific projects conducted by specialists to working within organizations to change modes of thinking, knowledge skills, and actions through helping individuals learn. As we discussed earlier, such changes are already happening in many organizations. Preskill and Boyle write, "We believe that ECB represents the next evolution of the evaluation profession and, as such, has the potential for transforming the field in ways only imagined" (Preskill & Boyle, 2008, p. 457).

But efforts at ECB and mainstreaming are in their infancy—or perhaps youth! We want to make readers aware of the issues and concepts, and we encourage you to read more. Just as Sanders' concepts for mainstreaming overlap with the definition given earlier for ECB, many definitions exist. The attention to the subject, and the diversity of approaches and concepts, is illustrated by a think tank on capacity building, funded by the National Science Foundation, at Utah State University's Consortium for Building Evaluation Capacity. Compton and Baizerman, after attending the session, observed that, "No two conceptions or definitions [of ECB] were the same" (2007, p. 118). They realize that this diversity is a natural stage in a practice-oriented field on a developing issue and that it is too early in the development of ECB to attempt to synthesize or control these different streams. Instead, as the AEA results show, ECB is being

practiced in many different ways, in many different settings. These definitions, models, and practices illustrate our point here: that evaluators have begun thinking about their impact, not just on programs, but on organizations and, in many cases, adapting their practices to encourage capacity building and mainstreaming of evaluation. Many readers of this book may not intend to become professional evaluators. Instead, they are managers or decision makers in their organization who want to make practical use of evaluations. Therefore, we include this discussion of ECB and mainstreaming to encourage the reader to pursue this role and to link evaluation to other change efforts in the organization and its culture.

Limitations to Mainstreaming Evaluation and Capacity Building

Mainstreaming and ECB are not without their critics. As Dahler-Larsen (2006) warns, evaluators do not gain their skills easily. Professional evaluators take years to build them and it can be misleading to imply that others, with other training and different responsibilities, can readily gain the same depth of skills. At the American Evaluation Association's 2001 convention on Sanders' theme of mainstreaming, Andersen and Schwandt questioned whether "the practice of establishing evaluation as a central (perhaps even an essential?) technology and ideology in social institutions, is an unqualified social good?" (2001, p. 1). They emphasized the need for a critical dialogue among those in the evaluation field and other stakeholders on the issue. Making evaluation a central, everyday component of organizations would constitute a fundamental change in organizations and has implications for society. What are the meanings and implications of those changes? Will evaluation be reduced to simple monitoring (Patton, 2001)? To issues that can only be accomplished by nonprofessionals? If so, what information, what values, will be lost? Evaluators have an ethical responsibility to include the perspectives of the full range of stakeholders, to take into account the public good, to use rigorous methods of systematic inquiry and to be competent in those methods, to ensure that the evaluation team is culturally competent, and to consider and pursue other important goals embodied in the *Standards* and the AEA *Guiding Principles* (American Evaluation Association, 2004; Joint Committee, 1994, 2010). Will our emphasis on organizational change—and with it the need to involve others in evaluation, build their competencies, and change our roles to lessen our control or leadership over evaluation—*reduce* our ability to uphold these ethical principles and standards? These questions, of course, cannot be answered now, but, as evaluators, we must recognize and discuss the strengths and limitations of ECB and mainstreaming as we would any intervention.

We see ECB and mainstreaming of evaluation as a goal that will improve organizations and their decisions and actions. Process use—changing nonevaluators' modes of thinking and decision making through their participation in evaluation—may be evaluation's greatest impact. We applaud and support this impact. But we also strive for instrumental use, and it is often achieved. Instrumental use, ironically, can

be increased through process use and ECB. That is, stakeholders who have learned about evaluation through participation are, one would think, more likely to seek and use results from evaluation in making decisions. (Now, we need to test that assumption!) We share the concerns of Dahler-Larsen and others. Not everyone will become an evaluator nor attain the skills of professional evaluators. Fortunately, many won't want to do so. They have their own professions. But, having them understand a little more about our approach to learning and judging cannot hurt and, we think, can help organizations change and better serve both their clients and society.

Major Concepts and Theories

1. Evaluations take place in different cultures, often ones that are new to the evaluator. To increase the validity and utility of the evaluation, and for ethical reasons, evaluators should become culturally competent in the setting of their evaluation.

2. As evaluations and evaluation requirements proliferate, evaluators and others have a role in helping to mainstream evaluation and to build an organization's capacity for conducting and using evaluations. These efforts are a major area of growth for evaluation today.

Discussion Questions

1. What is the culture of the course you are taking? Describe some of its characteristics. How would a person who has never been a student in a university course perceive this culture? Would he or she see it the same way that you do? If this person were an evaluator, how would his or her different perceptions affect the way the evaluation is conducted?

2. Do you think mainstreaming evaluation is possible? desirable?

Application Exercises

1. Read Fitzpatrick's interview with Katrina Bledsoe and her evaluation of the Fun with Books program in Trenton, New Jersey. (See *American Journal of Evaluation*, *28*(4), 522–535.) In it, she discusses cultural competence and her efforts to build it in her evaluation of this program. First, discuss and critique her methods for gaining cultural competence in the program. Then, discuss how you would obtain cultural competence to evaluate a program in your city that serves students or clients quite different from yourself or your experience. How would you build on Bledsoe's activities? What else would you do?

2. Read Jean King's chapter on her efforts at ECB in a school district and Bobby Milstein's efforts at ECB at the Centers for Disease Control. (Both are in *New Directions for Evaluation*, 2002, No. 93.) Compare and contrast their methods. Which do you favor and why? Which would be most appropriate for your own organization?

Case Studies

We recommend three interviews that touch on elements discussed in this chapter: *Evaluation in Action*, Chapters 8 (Jean King), 12 (Katrina Bledsoe), and 13 (Allan Wallis and Victor Dukay).

In Chapter 8, Jean King describes how she worked as an internal evaluator in a school district to build evaluation capacity both through the year and in a particular evaluation. Professor King takes a sabbatical from her faculty position to experience and learn from this role and has since focused her evaluation work on capacity building. This chapter is only available in the book.

In Chapter 12, Katrina Bledsoe, an African American from California, describes how she works to build cultural competence in a quite different African American community in Trenton, New Jersey. She also discusses how the different cultures of program staff, volunteers, and participants prompt them to have different views of the program's aims. The journal source for this chapter is Fitzpatrick, J.L., & Bledsoe, K. (2007). Evaluation of the Fun with Books Program: A dialogue with Katrina Bledsoe. *American Journal of Evaluation, 28,* 522–535.

In Chapter 13, Allan Wallis and Victor Dukay discuss an evaluation they conducted in rural Tanzania in Africa. The evaluation of an orphanage for children whose parents have died from AIDS leads them to involve Tanzanian researchers and laypeople in Tanzania to learn more about rural Tanzania and the culture of the place where the orphanage exists. This chapter is only available in the book.

Suggested Readings

Barnette, J. J., & Sanders, J. R. (Eds.). (2003). The mainstreaming of evaluation. *New Directions for Evaluation*, No. 99. San Francisco: Jossey-Bass.

Compton, D. W., Braizerman, M., & Stockdill, S. H. (Eds.). (2002). The art, craft, and science of evaluation capacity building. *New Directions for Evaluation*, No. 93. San Francisco: Jossey-Bass.

Fitzpatrick, J. L. (2008). Exemplars' choices: What do these cases tell us about practice? In J. Fitzpatrick, C. Christie, & M. M. Mark (Eds.), *Evaluation in action: Interviews with expert evaluators*, pp. 355–392. Thousand Oaks, CA: Sage.

Hood, S. L. (2001). Nobody knows my name: In praise of African American evaluators who were responsive. In J.C. Greene and T.A. Abma (Eds.), Responsive evaluation. *New Direction for Evaluation*, No. 92. San Francisco: Jossey-Bass.

Thompson-Robinson, M., Hopson, R., & SenGupta, S. (Eds.) (2004). In search of cultural competence in evaluation: Principles and practices. *New Directions for Evaluation*, No. 102. San Francisco, CA: Jossey-Bass.

10

A Comparative Analysis of Approaches

Orienting Questions

1. What are some cautions to keep in mind when considering alternative evaluation approaches?
2. Would you miss much if you ignored all but one approach, which you then used for all evaluations? What are some dangers in always using the same evaluation approach?
3. What has each of the alternative evaluation approaches contributed to the conceptualization of evaluation?

A Summary and Comparative Analysis of Evaluation Approaches

In Chapter 4, we presented a variety of ways to classify evaluation approaches. We included our schema, which organizes the proposed approaches into four categories: approaches to judge overall quality (expertise- and consumer-oriented), approaches oriented to characteristics of the program, decision-oriented approaches, and participant-oriented approaches. Together they represent the current major schools of thought about how to plan and conduct program evaluations. Collectively, Chapters 5 through 8 summarize the theoretical and conceptual underpinnings of most of today's program evaluations. It is therefore appropriate to ask how useful these frameworks are. The answer is, "Very useful indeed," as we shall discuss shortly, but first we feel compelled to offer several cautions.

Cautions about the Alternative Evaluation Approaches

Evaluation Approaches Are Distinct but May Be Mixed in Practice

In a young field, there is inevitably a good bit of conceptual floundering as notions are developed, circulated, tried out, revised, refined, and challenged again by new alternatives. Until a solid knowledge base begins to guide practice, any new field is likely to be guided by the training, experience, and views of its leaders. Some may argue that this is inappropriate, but it is also inevitable, since no new field or discipline is born full grown. Yet, it is also appropriate to ask how far the conceptions of leaders have led the field and in what direction.

Today, we have learned more and tried more in the field of evaluation than when we published our first edition of this book. Research on evaluation shows that almost all evaluators use mixed methods, both qualitative and quantitative; involve assorted stakeholders in the evaluation in some way; and consider ways to increase the use of their results (Christie, 2003; Fitzpatrick, 2004). Developers of evaluation approaches are much less likely to insist that their way is the only way, although they obviously believe that their own approach has advantages. Approaches that have lasted have changed with time and greater knowledge. Decision-oriented approaches consider stakeholders beyond managers; empowerment evaluation began to emphasize organizational learning and systems that empower staff and managers.

One area of potential confusion is our use of different words for evaluation approaches. Many well-known evaluators use the words "model" and "approach" interchangeably. Others write and study evaluation "theories" (Alkin & Christie, 2004; Chelimsky, 1998; Christie, 2003). In the past, we have argued that the word "approaches" is most appropriate for the nature of the subject matter we are considering, because they are too narrow to be theories and insufficiently tested to serve as models. But we too would like to incorporate other thinking from the field. Here, we are simply warning the reader that these words—approaches, models, and theories—are used interchangeably in the evaluation field. They all refer to views or systems that are used to guide the thinking about and conduct of evaluations.

One explanation for the decreased divergence and lessened rhetoric around evaluation approaches, models, or theories is the development and increased dissemination of two important documents to guide evaluation practice: (a) The Joint Committee on Educational Evaluation's *Program Evaluation Standards* that establish the criteria for good evaluation studies: utility, feasibility, propriety, accuracy, and accountability and (b) the American Evaluation Association's Guiding Principles that articulate expected ethical behaviors for evaluators in the categories of systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare. (See Chapter 3 for more on these *Standards* and Guiding Principles and Appendix A

for complete lists of each.) These documents and the discussions concerning them, their use in training, and the revisions of new editions, brought greater consensus to the field regarding the roles of evaluations, evaluators, stakeholders, and other users, and the public interest.

Discipleship to a Particular Evaluation Approach Is a Danger

> Into the street the Piper steps,
> Smiling first a little smile,
> As if he knew what magic slept
> In his quiet pipe the while; . . .
> —*Robert Browning,* "The Pied Piper of Hamelin"

Every evaluation approach described in this book has adherents who believe that a better evaluation will result from that orientation than from alternatives. Fair enough. We have no quarrel with those who follow particular persuasions as long as they do so intelligently, knowing when and where their preferred approach is applicable and when it may not be the best fit for the problem or the context.

What is troublesome, however, is that every evaluation approach also has some unthinking disciples who are convinced that a particular approach is right for every situation. There are evaluators who are CIPP loyalists, or unswerving adherents of theory-based evaluation, or those who hold the tenets of responsive evaluation as articles of faith. Many evaluators follow a chosen evaluation approach into battle without first making certain that the proposed strategy and tactics fit the terrain and will attain the desired outcomes of the campaign. Insisting that an outcome-based approach be used for an internal formative evaluation when the issues, not to mention the program, are vague and amorphous is as foolish as mounting a cavalry attack across a swamp.

Ideally, evaluation practitioners are knowledgeable enough about using these approaches as heuristic tools that they can select, from a variety of evaluation approaches, one that is appropriate for the program and its context and stakeholders instead of distorting the interests and needs of the evaluation's audience(s) to make them fit a preferred approach. For an example of a an appropriate use of a different model, see the interview with Fetterman in which he discusses departing from his empowerment approach in a particular evaluation because of the context and needs of the primary decision maker (Fitzpatrick & Fetterman, 2000).

Calls to Abandon Pluralism and Consolidate Evaluation Approaches into One Generic Model Are Still Unwise

Individuals new to evaluation can become frustrated with the different approaches. They may seem too abstract or irrelevant to the evaluation mandates of the funders or their supervisors. Some argue that we should develop one straightforward evaluation approach.

On the surface, such calls for synthesis are appealing because they address the desire of many practitioners and clients just to get to the point, namely, "Skip the academic discussions and just tell us how to do program evaluation!" Some textbooks do that, but they mislead the reader into thinking that evaluation is just a matter of collecting and analyzing data. We now recognize that evaluation is often played out in a highly political environment with stakeholders who have differing concerns and views and who may have designs on the evaluation, hoping it will support their side. Further, without the various approaches, in today's environment evaluators would think they always need to collect data on outcomes, attempt randomized control trials, and simply follow the dictates of their funding source or the regulations that govern their organization. Instead, the approaches help us recognize that important choices must be made and that, if these choices are slighted, the work and the cost of the evaluation will often be for naught. In other words, not taking the time to deliberate with stakeholders and others on the potential purposes of the evaluation and its intended or potential uses, failing to learn more about the program than its objectives or standards, and not listening to other stakeholders and considering whether and how to involve them, can result in an evaluation that meets the requirement, but is of little or no use to anyone. The approaches, and their differences, make us aware of the many important, nontechnical choices we must make to conduct a just, fair, and valid evaluation—one that will improve programs.

Another barrier to creating an approach that is a synthesis of existing approaches is that the ones described in the preceding chapters are based on widely divergent philosophical assumptions. Although some are compatible enough to be fruitfully combined, integrating all of them would be a philosophical impossibility, because key aspects of some approaches are directly incompatible with the central concerns of others. Transformative participatory approaches are based on quite different purposes and principles than decision-based approaches, for example. One is designed to transform society or at least shake up power arrangements. The other's more modest goal is to provide information, typically to those holding power, to make decisions. No synthesis is possible there. Similarly, focusing on defining the purpose is quite different from focusing on program outcomes or theory or, for that matter, focusing on the qualities identified by expert connoisseurs. These, too, are likely to be incompatible in most programs.

Moving toward one omnibus model at this time could also bring premature closure to expansion and refinement within the field. Just as there are many different interpretations of evaluation capacity building, our concepts of evaluation are still too untried and our empirical base too weak for us to know which notions should be preserved and which discarded. It would seem far better to tolerate our contradictory and confusing welter of ideas and make of them what we can than to hammer them into a unified but impoverished concept of evaluation. In writing more broadly on the conduct of inquiry and scientific research, Kaplan wrote, "The dangers are not in working with models, but in working with too few, and those too much alike, and above all, in belittling any efforts to work with anything else" (Kaplan, 1964, p. 293). Just because we can synthesize does not mean that

we should. As Kaplan puts it, consolidation would impose a premature closure on our ideas and thus limit

> our awareness of unexplored possibilities of conceptualization. We tinker with the model when we might be better occupied with the subject-matter itself . . . incorporating it in a model does not automatically give such knowledge scientific status. The maturity of our ideas is usually a matter of slow growth, which cannot be forced. . . . Closure is premature if it lays down the lines for our thinking to follow when we do not know enough to say even whether one direction or another is the more promising. (p. 279)

A final concern has to do with the huge diversity of contexts—different countries; different political systems; different funder, citizen, parent, client, and student expectations—in which evaluation is conducted. Because evaluation contexts are so different, it is difficult to conceive of any one model that would be relevant to all. For all their imperfections, diverse frameworks offer a richness of perspectives and serve as heuristics, especially if one uses evaluation approaches eclectically (when philosophical compatibility permits), as we propose.

The Choice of Evaluation Approach Is Not Empirically Based

If one accepts our view that it is useful to have a variety of evaluation approaches, the next logical question is how will one know which approach is best for a given situation? That question is devilishly difficult to answer because of one simple fact: There is almost no research to guide one's choice. Although research on the practice of evaluation is increasing (See Cousins, Donohue, & Bloom [1996]; Cullen, 2009; Christie, 2003; Fitzpatrick, 2004), it may not be the best way for a practicing evaluator to select an approach, although it can be helpful in illustrating what others do. Today's research suggests that people, even theorists, apply the approaches eclectically, adapting to the needs of the decision-makers and stakeholders, the program, and the context in which it operates.

So the choice is made based on knowledge of the available approaches and their methods and reasoning about which approach best fits the evaluation being done. We encourage you to read the interviews and cases we have listed in many of the chapters to learn more about how exemplary evaluators make choices in particular contexts and their reflections on those choices. Evaluation is an art and a practice. Making choices about approaches cannot be based on empirical evidence, because studies could not cover all the characteristics of each locality that evaluators face. Instead, the evaluator must take time to explore the environment and to consider which approaches fit best. In the next section of this book, we will discuss methods you can use during the planning phase to learn more about the program, the context, and the stakeholders, and to make appropriate choices.

Contributions of the Alternative Evaluation Approaches

If, as research has found, many practitioners do not follow a single approach and theorists "mix it up," using elements of different approaches, what is the worth of the approaches anyway? Considerable worth, actually. An evaluator may never use Scriven's (1972) goal-free method of evaluation, but the concept reminds us to think beyond the goals, to look for unintended side effects—both good and bad—and to be aware of how our knowledge of goals constrains us, like blinders on a horse, thereby inhibiting us from noticing other outcomes or impacts. Individuals may spend years as evaluators and never once use Stufflebeam's (1971) CIPP model of evaluation, but many have used his concept of program stages and the different information needs that may arise at each stage. Further, the model has made them recognize that evaluation can effectively provide information for decisions before the program even begins. Similarly, most of the evaluation approaches summarized in the previous chapters influence the practice of evaluation in important ways.

Thinking back to evaluations we have done, our colleagues' work was used in this way in almost every study. As one of us noted in earlier editions of this textbook:

> Although I have developed some preferences of my own in doing evaluations, probably 75 percent of what I do is application of what I have distilled from others' ideas. Doubtlessly, all who have been repeatedly exposed to the evaluation literature have absorbed much 'through the pores,' as it were, and now reapply it without cognizance of its source. Although few of us may conduct our evaluations in strict adherence to any 'model' of evaluation, few of us conduct evaluations which are not enormously influenced by the impact of our colleagues' thinking on our own preferences and actions. (Worthen, 1977, p. 12)

The alternative concepts about how evaluation should be conducted—the accompanying sets of categories, lists of things to think about, descriptions of different strategies, and exhortations to heed—influence the practice of program evaluation in sometimes subtle, sometimes direct, but always significant ways. Some evaluation designs adopt or adapt proposed approaches. Many evaluators, however, conduct evaluations without strict adherence (or even purposeful attention) to any model. Instead, they draw unconsciously on the philosophy, plans, and procedures they have internalized through exposure to the literature. So, the value of the alternative approaches lies in their capacity to help us think, to present and provoke new ideas and techniques, and to serve as mental checklists of things we ought to consider, remember, or worry about. Their heuristic value is very high; their prescriptive value seems much less.

Comparative Analysis of Characteristics of Alternative Evaluation Approaches

So many new concepts have been presented in Chapters 5 through 8 that the reader might be feeling challenged to assimilate all of it. The matrix in Table 10.1—a comparative analysis of the characteristics, strengths, and limitations of the four approaches—should help. The aspects of each approach that we have chosen to highlight are as follows:

1. Proponents—Individuals who have written about the approach
2. Purpose of evaluation—The intended focus of evaluation proposed by writers advocating each particular approach or the purposes that may be inferred from their writings
3. Distinguishing characteristics—Key factors or concepts associated with each approach
4. Benefits—Strengths that may be attributed to each approach and reasons why one might want to use it (what it can do for you)
5. Limitations—Risks associated with the use of each approach (what it can do to you)

TABLE 10.1 *Comparative Analysis of Alternative Evaluation Approaches*

	Expertise-Oriented	*Consumer-Oriented*
Some proponents	Eisner Accreditation groups	Scriven Consumers Union
Focus of evaluation	Providing professional judgments of quality	Judging quality of products to aid decisions about purchases
Distinguishing characteristics	Basing judgments on individual knowledge and experience; in some cases, use of consensus standards, team/site visitations	Using criterion checklists to analyze products; product testing; informing consumers
Uses	Criticism; self-study; accreditation; blue-ribbon panels	Consumer reports; product development; selection of products for dissemination
Contributions to evaluation	Subjective criticism as a form of disciplined inquiry; self-study with outside verification; standards	Articulating criteria, using checklists; formative-summative purposes; bias control
Strengths	Applicable to many areas; efficiency (ease of implementation, timing)	Emphasis on consumer information needs; developing checklists; influence on product developers; concern with cost-effectiveness and utility
Limitations	Reliability; replicability; vulnerability to personal bias; scarcity of supporting documentation to support conclusions; open to conflict of interest.	Lack of sponsors or funders; not open to debate or cross-examination

(continued)

TABLE 10.1 ***Comparative Analysis of Alternative Evaluation Approaches (Continued)***

	Program-Oriented	***Decision-Oriented***
Some proponents	Tyler Provus Weiss Chen Bickman Donaldson	Stufflebeam Alkin Provus Patton Wholey
Focus of evaluation	Determining the extent to which program objectives or key elements in the program theory are delivered or achieved	Providing useful information to aid in making decisions
Distinguishing characteristics	Focus on identifying and describing key elements of program and why it should work; uses dialogue with program people to develop theory; builds on developers' theories and related research; can be more quantitative and causal	Serving rational decision making; evaluating at all stages of program development; working with managers to increase use
Past uses	Program planning and development; adding to knowledge; evaluating program outcomes and links to those outcomes	Program development; organization's management systems; program planning; accountability
Contributions to evaluation	Considering program theory and link between program activities and outcomes; using research literature as a source of ideas and methods	Link evaluation to decisions; work closely with managers to identify decisions and to learn about the context in which decisions are made; provide information in an ongoing manner through information systems
Strengths	Prompts link to program creators and to research literature; helps explain program outcomes, avoids black box (unknowns about outcomes); emphasis on explanation of program outcomes	Potentially comprehensive; sensitive to information needs of those in a leadership position or who will use evaluation; systematic approach to evaluation; use of evaluation throughout the process of program development
Limitations	May be too concerned with research, less with stakeholders; possible overemphasis on outcomes	Assumption of orderliness, rationality, and predictability in decision making; focus on the concerns of managers or leaders

TABLE 10.1 ***Continued***

	Participant-Oriented
Some proponents	Stake Guba and Lincoln Fetterman Cousins and Earl House and Howe
Focus of evaluation	Involving many stakeholders or a few stakeholders in depth in the evaluation; understanding and portraying the complexities of programmatic activity; empowering stakeholders; pursuing social justice
Distinguishing characteristics	Reflecting multiple realities; intensive involvement with stakeholders and understanding of context; greater use of qualitative methods; greater focus on formative evaluation and organizational learning
Uses	Smaller evaluations; focus can be on process use and organizational learning; transformative evaluations used internationally
Contributions to evaluation	Emergent evaluation designs; using stakeholders to improve use, evaluator's understanding, democracy, dialogue; importance of the particulars of the program; understanding and attention to context
Strengths	Pluralistic; focus on description and judgment; emphasis on understanding and use; recognition and pursuit of different types of use, including organizational and individual learning; focus on the particulars of a program
Limitations	Potentially high labor intensity and cost; potential for stakeholders with less knowledge of evaluation to lead study inappropriately; less generalizable; less replicable

Eclectic Uses of the Alternative Evaluation Approaches

The purpose of the foregoing comparative analysis is to provide key information on the strengths, limitations, and primary uses of each approach. The information in Table 10.1 is not intended to imply that any one approach is best; rather, it is our contention that each approach can be useful. The challenge is to determine which approach (or combination of concepts from different approaches) is most relevant to the task at hand.

Perhaps the experience of one of the authors in attempting to answer a student's question in a graduate evaluation seminar will help make the point.

We were conducting a several-week-long examination of various authors' evaluation approaches and how each might be applied to do an evaluation, when one student asked, "What approach do you usually use?" I pointed out that I did not believe there was one best approach, that each has its strengths, and that I simply used whichever approach was most appropriate to the situation at hand.

"How do you know which one is most appropriate?" she queried. I talked about things like looking at the purpose of the evaluation, the kind of decision needed, limitations of the approach, and so on, and concluded that a lot of it was in experience and, although a little tough at first, they would all get the hang of it once they had done a few evaluations.

"Maybe it would help," she stated, "if you could give us a few examples of where you've used one of the approaches and then show us why you picked it."

That seemed like a very useful suggestion, so I began to sort through my mental files to find the very best examples of where I had used one of the evaluation approaches. Then I began to sort to find any examples of where I had used one of the approaches. I discarded evaluation after evaluation because I really had not used the approach, whatever it was, fully. There were truncated CIPP evaluations, because I seldom seemed to be called on early enough to do much with context or input evaluations. There were applications of responsive evaluation through involving stakeholders in different ways. Each was incomplete as an example of use of the approaches, and I struggled for more pure examples to offer.

Finally, I remembered using Stake's "countenance" framework in its entirety in evaluating an administrators' training program. That one was memorable because it had been a class project that two students and I had taken on so they could get the experience. That one brought others to mind and before long I was able to give examples of using several of the frameworks in the way they were intended to be used. The intriguing realization was that every one of those examples came from class projects conducted jointly with students, when I had intentionally adhered to the models to demonstrate their features. I could not recall a single "lone-wolf" evaluation of my own for which I had consciously selected any single approach to guide the study. Instead, for several years I had been designing each evaluation de novo, pulling pieces of different frameworks in as they seemed relevant. Certain features of some models I used frequently, others seldom or never. That realization seemed worth sharing, although in the process I felt a twinge of disloyalty toward some of my esteemed colleagues and friends for never really using their frameworks completely in my work. The class was slightly taken aback at first by my heretical revelation, but they seemed comforted when I pointed out that there were distinct advantages in eclecticism, because one was free to choose the best from diverse sources, systems, or styles. Warming to the idea, I argued that one could choose the best features of each approach and weave them into a stronger overall approach—really a classic bit of cake-having and cake-eating.

We talked for the remainder of the class about why each evaluation required a somewhat different mix of ingredients, how synthesis and eclecticism were not identical, and why an eclectic approach could be useful. (Worthen, 1977, pp. 2–5)

The authors of this text are all self-confessed eclectics in our evaluation work, choosing and combining concepts from the evaluation approaches to fit the particular situation, using pieces of various evaluation approaches as they seem appropriate.

In very few instances have we adhered to any particular model of evaluation. Rather, we find we can ensure a better fit by snipping and sewing together bits and pieces of the more traditional ready-made approaches and even weaving a bit of homespun, if necessary, rather than by pulling any existing approach off the shelf. Tailoring works.

Obviously, eclecticism has its limitations—after all, it has been derided as the discipline of undisciplined minds. One cannot suggest that we develop an "eclectic model" of evaluation, for that would be an oxymoron. The uninformed could perform egregious errors in the name of eclecticism, such as proposing that a program's objectives be evaluated as a first step in conducting a goal-free evaluation or laying out a preordinate design for a responsive evaluation. Assuming that one avoids mixing evaluation's philosophically incompatible oil and water, the eclectic use of the writings presented in the preceding chapters has far more potential advantages than disadvantages, whether that eclecticism means combining alternative approaches or selectively combining the methods and techniques inherent within those approaches.

Eclecticism is more common in education, however, than in some other areas. This is partly because education has been the primary field in which different approaches have been developed. Other fields, such as sociology, criminal justice, and mental health, have erred in not considering those approaches for evaluation. By failing to do so, evaluators in those fields have often failed to consider sufficiently the critical components of their evaluation, such as audiences, purposes, and uses. The evaluations in those fields have remained more applied research than evaluation and have been more summative than formative. Much of evaluation's potential lies in the scope of strategies it can employ and in the possibility of selectively combining those approaches. Narrow, rigid adherence to single approaches must give way to more mature, sophisticated evaluations that welcome diversity. Admittedly, this will be a challenging task, but that does not lessen its importance.

Drawing Practical Implications from the Alternative Evaluation Approaches

All the evaluation approaches we have presented have something to contribute to the practicing evaluator. They may be used heuristically to generate questions or to uncover issues. The literature contains many useful conceptual, methodological, political, communicative, and administrative guidelines. Finally, the approaches offer powerful tools that evaluators may use or adapt in their work.

Later in this book, we will look at practical guidelines for planning and conducting evaluations. Many of these guidelines have been developed as part of a particular approach to evaluation. Fortunately, however, they are generalizable—usable whenever and wherever needed. Just as a skilled carpenter will not use only a hammer to build a fine house, so a skilled evaluator will not depend solely on one approach to plan and conduct a high-quality evaluation.

In the next section of this book, we will look at practical uses of the tools that evaluation practitioners, theorists, and methodologists have generated. But first, pause a moment to apply what you have learned in this chapter.

Major Concepts and Theories

1. The different approaches to evaluation are natural and should not be consolidated or synthesized. The differences they reflect mirror the many different purposes and contexts for evaluation and prompt the evaluator to actively consider his or her choices in conducting an evaluation.

2. Evaluators determine which approach(es) to employ in a given situation based on the context in which the evaluation is conducted, including the stage of the program, the needs of the stakeholders, and the culture of the organization.

3. Often, evaluators will not adhere to one specific approach, but instead will opt for a combination of several approaches in a more eclectic approach to evaluation.

Discussion Questions

1. Why is it important to learn about the different evaluation approaches?

2. How would you go about selecting an evaluation approach to employ in your school or organization? What factors would be important? Which approach might you select or would you adopt parts of several approaches? If the latter, what would you use?

3. Is there one evaluation approach that you particularly like or feel more comfortable with than another? Why?

4. Should we attempt to synthesize the different evaluation approaches into one? What would be the advantages and disadvantages of doing so?

Application Exercises

1. Identify five evaluation studies in a journal of interest to you or, better yet, collect in-house reports on five evaluation studies. These might be from your own welfare agency; from your school or university; from a city, county, state, or federal office; or from a nonprofit agency. After reading the report, discuss what approach the author used. Is it eclectic, or does it follow one model predominantly? What elements of each approach seem to be most useful in guiding the authors in identifying their purpose, audiences, data collection methods, and presentation of results? Using each of the four evaluation approaches discussed in Chapters 5 through 8, discuss how that approach might lead you to proceed differently with the evaluation. Do you see any combination of approaches that would be particularly useful in evaluating this program?

2. Following is a list of evaluation purposes. Which approach would you use in each of these examples? Why? What would be the advantages and disadvantages of this approach in each setting?

a. Determining whether to continue a welfare-to-work program designed to get full-time, long-term employment for welfare recipients

b. Describing the implementation of a distance-learning education program for college students

c. Making recommendations for the improvement of an anti-bullying program for middle-school students

d. Determining whether reading levels of first-graders at the end of the year are appropriate.

3. Adams Elementary School has started a volunteer program in which parents are encouraged to help out in the classroom. The goal of the program is not only to provide the teacher with assistance, but also to get parents more involved in the school and their children's education. The principal hopes to boost the learning of the students who are achieving below grade level by getting their parents more involved in their children's education through volunteer efforts in the classroom. Contrast using a program-oriented, decision-oriented, and participant-oriented approach.

Case Studies

Rather than recommend an individual case study for this synthesis chapter, we recommend that you look back over the case studies you have read and, for each, identify which evaluation approach or approaches they are using. Consider the extent to which they are true to one approach or make use of different approaches.

Suggested Readings

Christie, C. A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. In C. A. Christie (Ed.), *The practice-theory relationship in evaluation.* New Directions for Evaluation, No. 97, 7–36. San Francisco: Jossey-Bass.

Fitzpatrick, J. L. (2008). Exemplars' choices: What do these cases tell us about practice? In J. Fitzpatrick, C. Christie, & M. M. Mark (Eds.), *Evaluation in action: Interviews with expert evaluators*, pp. 355–392. Thousand Oaks, CA: Sage.

Stufflebeam, D. L. (2001). Evaluation models. *New Directions for Evaluation*, No. 89. San Francisco, CA: Jossey-Bass.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications.* Somerset, NJ: John Wiley.

This page intentionally left blank

Part III

Practical Guidelines for Planning Evaluations

In Part One, we introduced the reader to evaluation and its history and to some of the critical political and ethical issues that surround and define evaluation. In Part Two, we examined factors that led to alternative conceptions of evaluation, summarized the key characteristics and the strengths and weaknesses of the four general evaluation approaches that are most influential today, and argued for thoughtful use of those approaches, including eclectic combinations.

This brings us to the heart of this book: practical guidelines. In this part we begin to provide guidelines that will be helpful to evaluators, regardless of which evaluation approach or combination of approaches they might elect to use. In Part Three, we present guidelines for clarifying, focusing, and planning evaluation efforts. Then, in Part Four, we will present guidelines for conducting and using evaluations.

We begin Part Three by examining in Chapter 11 the reasons that lead to initiation of program evaluations, considerations in deciding when to evaluate (*always* is a common but incorrect answer), and how to determine who should conduct the evaluation. In Chapter 12, we discuss the importance of the evaluator's understanding the setting and context in which the evaluation will take place, as well as the importance of accurately describing that which is to be evaluated. Two crucial steps in evaluation planning—identifying and selecting evaluative questions and criteria, and planning the information collection, analysis, and interpretation—are examined in detail in Chapters 13 and 14. Also included in Chapter 14 are a few guidelines for developing management plans for evaluation studies, stressing the importance of establishing evaluation agreements.

The focus of these chapters is decidedly practical. Although we will continue to quote or reference other sources, the chapters are not intended as scholarly reviews of the content covered. Were such reviews to be included, several of these

chapters could each fill a textbook. Our intent is only to introduce enough information to give both the evaluator and the user of the evaluation (1) an awareness of how to proceed and (2) direction to more detailed coverage of many (especially technical) topics in other textbooks. Experience and further study will have to teach the rest.

11

Clarifying the Evaluation Request and Responsibilities

Orienting Questions

1. Suppose you received a telephone call from a potential client asking if you would do an evaluation. What are some of the first questions you would ask?
2. Are there times you would decline a request for evaluation? If so, under what conditions?
3. How can an evaluability assessment help determine whether an evaluation will be productive?
4. What are some advantages and disadvantages in having an evaluation conducted by an external evaluator? By an internal evaluator?
5. What criteria would you use to select an external evaluator?

In the preceding chapters, we discussed evaluation's promise for improving programs. The potential and promise of evaluation may create the impression that it is always appropriate to evaluate and that every facet of every program should be evaluated.

Such is not the case. The temptation to evaluate everything may be compelling in an idealistic sense, but it ignores many practical realities. In this chapter we discuss how the evaluator can better understand the origin of a proposed evaluation and judge whether or not the study would be appropriate.

To clarify the discussion, we need to differentiate among several groups or individuals who affect or are affected by an evaluation study: sponsors, clients, stakeholders, and audiences.

An evaluation's sponsor is the agency or individual that either requests the evaluation or provides necessary fiscal resources for its conduct, or both. Sponsors may or may not actually select the evaluator or be involved in shaping the study, but they often define the purposes of the evaluation and may specify particular areas that the evaluation should address or ways in which data should be collected. In other cases, the sponsor may delegate that authority to the client. The sponsor may be a funding agency or a federal or state department that oversees or regulates the activities of the organization that delivers the program.

The client is the specific agency or individual who requests the evaluation. That is, the client seeks an evaluator—internal or external—to conduct the evaluation and typically meets frequently with that evaluator as the evaluation proceeds. In some instances, the sponsor and client are the same, but not always. For example, in an evaluation of a domestic violence treatment program operated by a nonprofit agency, the agency (client) requests and arranges for the study, but the requirement and the funding may both originate with a foundation that funds the program and is, therefore, the sponsor. In contrast, the sponsor and the client are the same if the program to be evaluated is a drop-out prevention program for district high schools that is funded by the school district, and the person requesting the evaluation is a central office administrator who oversees secondary programs.

As we discussed in Chapter 1, stakeholders consist of many groups, but essentially include anyone who has a stake in the program to be evaluated or in the evaluation's results. Sponsors and clients are both stakeholders, but so are program managers and staff, the recipients of program services and their families, other agencies affiliated with the program, interest groups concerned with the program, elected officials, and the public at large. It is wise to consider all the potential stakeholders in a program when planning the evaluation. Each group may have a different picture of the program and different expectations of the program and the evaluation.

Audiences include individuals, groups, and agencies who have an interest in the evaluation and receive its results. Sponsors and clients are usually the primary audiences and occasionally are the only audiences. Generally, though, an evaluation's audiences will include many, if not all, stakeholders. Audiences can also extend beyond stakeholders. They can include people or agencies who fund or manage similar programs in other places or who serve similar populations and are looking for effective programs.

Understanding the Reasons for Initiating the Evaluation

It is important to understand what prompts an evaluation. Indeed, determining and understanding the purpose of the evaluation is probably the most important job the evaluation sponsor or client will have in the course of an evaluation. If some problem prompted the decision to evaluate, or if some stakeholder or sponsor has demanded an evaluation, the evaluator should know about it. In many cases today, an evaluation is conducted in response to a mandate from a funding

source that is concerned about being accountable to a board or the public about programs it has funded. Presumably, the decision to evaluate was prompted by someone's need to know something. Whose need? What does that policymaker, manager, stakeholder, or agency want to know? Why? How will they use the results? The evaluator's first questions should begin to identify these reasons.

Sometimes the evaluation client can answer such questions directly and clearly. Unfortunately, that is not always the case. As evaluation has become popular today, often evaluations are undertaken or mandated for few clear reasons other than that evaluation is a good thing or that programs should be accountable. Of course, the evaluator's task is made more difficult when the client has no clear idea about what the evaluation should accomplish. It is not uncommon to find that clients or sponsors are unsophisticated about evaluation procedures and have not thought deeply about the purposes of the evaluation and the variety of questions it could answer or issues it could address. Worse yet, they may think that all evaluations automatically address outcomes or impacts and may insist that all evaluations address the same issues regardless of the stage of the program, the decisions they or others face, or the information needs of other stakeholders.

Frequently, the purpose of the evaluation is not clear until the evaluator has carefully read the relevant materials, observed the evaluation object, and probed the aspirations and expectations of stakeholders through significant dialogue.

Such probing is necessary to clarify purposes and possible directions. When sponsors or clients are already clear about what they hope to obtain, it is crucial for evaluators to understand their motivations. They can often do so by exploring—with whomever is requesting the evaluation and other stakeholders—such questions as the following:

1. *Purpose.* Why is this evaluation being requested? What is its purpose? What questions will it answer?

2. *Users and Use.* To what use will the evaluation findings be put? By whom? What others should be informed of the evaluation results?

3. *The Program.* What is to be evaluated? What does it include? What does it exclude? When and where does it operate? Who is the intended client for the program? What are the goals and objectives of the program? What problem or issue is the program intended to address? Why was it initiated? Who was involved in its planning? What prompted the selection of this strategy or intervention? Who is in charge of the program? Who delivers it? What are their skills and training? Has it ever been evaluated before? What data exist on it?

4. *Program logic or theory.* What are the essential program activities? How do they lead to the intended goals and objectives? What is the program theory or logic model? What have different stakeholders observed happening as a result of the program?

5. *Resources and timeframe.* How much time and money are available for the evaluation? Who is available to help with it? What is the timeframe for it? When is final information needed? Are there requirements that must be met for interim reports?

6. *Relevant contextual issues.* What is the political climate and context surrounding the evaluation? Who are the most concerned stakeholders? What individuals or groups might benefit from a positive evaluation? Who might benefit from a negative one? Will any political factors and forces preclude a meaningful and fair evaluation?

The foregoing questions are examples, and evaluators might subtract some or add others. What is important is that, through careful questioning, listening, and dialogue, the evaluator comes to understand the purpose for the evaluation and learns more about the context in which the program operates. Not all purposes are equally valid. By listening closely to the client's reasons for initiating the evaluation and talking with other stakeholders to determine their information needs and expectations for the study, the evaluator can learn much that will help ensure that the evaluation is appropriately targeted and useful.

The evaluator can also take a proactive role during this phase by suggesting other reasons for evaluating that may prove even more productive (Fitzpatrick, 1989). This strategy is particularly useful when the stakeholders are new to evaluation and unsure of their needs. Sometimes clients assume they must follow the sponsor's guidelines when a little dialogue with the sponsor might reveal more flexibility and open up avenues that will be more useful for the client in improving the program. Some clients or sponsors may assume that evaluations should only measure whether objectives are achieved or describe program outputs, outcomes, or impacts when, in fact, other critical information needs exist that could be served by evaluation. (For example, programs in their early stages often benefit from describing what is happening in the program, whether its activities are being delivered as planned, and whether adaptations are required.) Other clients may want to rush into data collection, seeing the evaluator's role as "helping us with a survey" or "analyzing some test scores." They are unfamiliar with the critical planning phase and how the evaluator can help them focus the evaluation to determine what they want to know. This phase begins the important two-way communication process essential to evaluation, in which the evaluator learns as much as he or she can about the program through careful questioning, observing, and listening and, at the same time, educates the sponsor, client, or other stakeholders about what evaluation can do.

In the early days of evaluation, Cronbach emphasized the importance of the educative role of the evaluator in helping the client determine the directions of the evaluation. Others emphasize that role today (Fitzpatrick & Bickman, 2002; Schwandt, 2008). Cronbach and his colleagues write that "the evaluator, holding the mirror up to events, is an educator. . . . The evaluator settles for too little if he simply gives the best answers he can to simple and one-sided questions from his clients. He is neglecting ways in which he could lead the clients to an ultimately more productive understanding" (1980, pp. 160–161). Therefore, before proceeding with the evaluation, the evaluator must spend a significant period of time learning about the program, its stakeholders, the decision-making process, and the culture of the organization to accurately determine the purpose of the study.

Direct Informational Uses of Evaluation

Evaluation is intended to enhance our understanding of the value of whatever is evaluated. Yet, as we noted at the beginning of this text, evaluation has many different uses. Examples of some of the informational uses of evaluation by policy makers, program managers, and program staff include:

1. Determining whether sufficient need exists to initiate a program and describing the target audience
2. Assisting in program planning by identifying potential program models and activities that might be conducted to achieve certain goals
3. Describing program implementation and identifying whether changes from the program model have occurred
4. Examining whether certain program goals or objectives are being achieved at the desired levels
5. Judging the overall value of a program and its relative value and cost compared with competing programs

Each of these five uses may be directed to an entire program or to one or more of the smaller components of a program. The first two uses are frequently part of planning and needs assessment (Altschuld, 2009; Witkin & Altschuld, 1995). These tasks generally take place during the early stages of a program, but they may occur at any stage in which program changes are being considered. The third use is often described as a monitoring or process evaluation. The fourth one can be characterized as an outcome or impact study. The final use is achieved through conducting cost-effectiveness or cost-benefit studies. All of these studies serve legitimate uses for evaluation because each one serves an important, informational use: enhancing our understanding of the value of the program.

Noninformational Uses of Evaluation

In addition to the direct informational uses described in the previous section, evaluation also has important noninformational uses. Cronbach and his colleagues (1980) first noted this in arguing that the very incorporation of evaluation into a system makes a difference. They conclude that "the visibility of the evaluation mechanism changes behavior" (p. 159), citing as an analogy how drivers' observance of speed limits is affected by police officers patrolling the highways in plainly marked patrol cars. They also suggest that the existence of evaluation may help convince stakeholders that the system is responsive, not impervious, to their feedback.

As the approaches in Part Two indicated, evaluations have many other impacts. One important use is its role in educating others, not simply about the program being evaluated, but also about alternative means for decision making. Smith (1989) writes that one of the most important benefits of evaluability assessment, a method of determining whether the program is ready for evaluation, is improving the skills of program staff in developing and planning programs.

Through participating in an extensive series of structured discussions to develop the program model, program staff gain skills that can be used in the next program they develop, even if evaluation is not involved. These kinds of changes represent organizational learning, where evaluations influence how the organization, its managers, and staff make decisions (Preskill & Torres, 2000).

Evaluation can also educate stakeholders, empowering them to become active in evaluations, by helping them gain skills in questioning and learning about programs in which they have a stake. Fetterman's empowerment evaluation approach, Cousins and Earl's practical participative evaluation approach, and Patton's utilization-focused evaluation all involve evaluators working closely with stakeholders and thereby increasing their skills in evaluation and data-based decision making.

Others (House & Howe, 1999) propose that evaluation can be useful in helping gain social equity and equality for stakeholders with less power. They note that evaluators often inform the stakeholders who have more power, such as policymakers, school boards, and legislators, because they are often the ones with the resources to commission an evaluation. A noninformational use of the deliberative democratic approach proposed by House and Howe is to help improve democracy by including other, less powerful stakeholders in the questions and discussions that emerge in evaluation.

However, in her seminal treatise on evaluation, Weiss (1972) noted that evaluation also has several undesirable, noninformational uses that are seldom acknowledged. Following are some of the more covert, nefarious, and patently political uses that she cites:

> *Postponement.* The decision maker may be looking for ways to delay a decision. Instead of resorting to the usual ploy of appointing a committee and waiting for its report, he can commission an evaluation study, which takes even longer.
>
> *Ducking responsibility.* . . . There are cases in which administrators know what the decision will be even before they call in the evaluators, but want to cloak it in . . . legitimate trappings. . . .
>
> *Public relations.* . . . The administrator believes that he has a highly successful program and looks for a way to make it visible. . . . The program administrator's motives are not, of course, necessarily crooked or selfish. Often, there is a need to justify the program to the people who pay the bills, and he is seeking support for a concept and a project in which he believes. . . .
>
> *Fulfilling grant requirements.* . . . Many federal grants . . . are tagged with an evaluation requirement. . . . [T]he operators of a project . . . tend to neglect the evaluation [and] . . . see it mainly as a ritual designed to placate the funding bodies, without any real usefulness to them.
>
> Evaluation, then, is a rational enterprise sometimes undertaken for nonrational, or at least noninformational, reasons. (pp. 11–12)

More recent work by Worthen (1995) suggests, however, that such noninformational uses may be more common in federal or national evaluations than in those administered at the state or local level. In an analysis of 108 evaluations, Worthen found that more than two-thirds of the evaluations of state and local programs served informational purposes, whereas only 15 percent of those con-

ducted at the federal level served such purposes. Although these results are based on a sample of studies conducted by only one institute, the Western Institute for Research and Evaluation (WIRE), and the number of national programs sampled is relatively small, the results do fit rather well with our collective experiences in other evaluations. If one can assume that the political tides run stronger in national programs than at lower levels, then these results may be attributable to the impact political forces have on evaluation, as we discussed in Chapter 3.

Conditions under Which Evaluation Studies Are Inappropriate

Except for some uses cited by Weiss, the foregoing examples all represent appropriate uses of evaluation studies. But evaluations are not always used appropriately. Smith (1998) has outlined several reasons for declining an evaluation contract. He groups them into two broad categories: (1) when the evaluation could harm the field of evaluation, or (2) when it would fail to support the social good. These problems may arise when it is likely that the ultimate quality of the evaluation will be questionable, major clients would be alienated or misled concerning what evaluation can do, resources will be inadequate, or ethical principles would be violated. Building on Smith's typology, we will outline several circumstances in which evaluations are, at best, of dubious value.

Evaluation Would Produce Trivial Information

Heretical as this may sound to some, sometimes a program simply lacks sufficient impact to warrant the expense of formal evaluation. Some programs are one-time efforts with no potential for continuation. Some are provided at such low cost to so few people that the need for more than informal evaluation is unlikely. Evaluability assessment or other evaluation planning activities may indicate that the program's theory or model is inadequate to achieve the desired impact. In other words, program activities simply have insufficient connections to the program goals or are too weak, due to duration or intensity, to achieve desired outcomes. Needs assessments or formative evaluations may be of use in improving the program if the emphasis in the evaluation is formative. However, summative or outcome evaluations are probably not worth the cost unless there is a need to demonstrate failure. Common sense must dictate when a program has enough impact to warrant a formal evaluation of its effectiveness.

Evaluation Results Will Not Be Used

Too often the professed need for an evaluation is merely an unreasoned assumption that every program must be evaluated. Evaluation is of dubious value unless there is commitment by someone to use the results. Given the scarcity of

evaluation resources (both financial and human) and the demand for evaluation information to inform important decisions, it seems a questionable investment at present.

Sometimes there are important decisions or choices to be made, but it is clear that they will be made for reasons unrelated to evaluative data. A program may, for instance, have sufficient political appeal or public support that administrators are simply unwilling to discontinue or change it drastically, no matter what problems an evaluation study may reveal. For example, Drug Abuse Resistance Education programs, better known as DARE, have enjoyed wide public support in spite of the failure of repeated, rigorous evaluations to find an effect on subsequent drug use (Lyman, Milich, Zimmerman, Novak, Logan, & Martin, 1999; Rosenbaum & Hanson, 1998; St. Pierre & Kaltreider, 2004). In this case, evaluation can play no meaningful role. Evaluators should avoid meaningless, ritualistic evaluations or pro forma exercises in which evaluation only appears to justify decisions actually made for personal or political reasons.

Of course, such dubious (and, one hopes, rare) motives are not always apparent. One of the most frustrating situations the evaluator will confront is to learn, after the evaluation has been completed, that the client or sponsor was not really open to information that contradicted preconceived notions. If the evaluator learns during the evaluation that certain conclusions are inevitable, it would be best to find ways to truncate the evaluation sham at the earliest opportunity.

Evaluation Cannot Yield Useful, Valid Information

Sometimes, despite an important pending decision, it appears highly unlikely that an evaluation study will produce any relevant information. For example, consider a decision about whether to continue a school dropout-prevention program. Here information about the program's effects on dropout rates, graduation percentages, and so forth would be relevant. But what if the program only started one month before the school board must make its decision? The probability of obtaining dependable information (even predictive information) about the program's effectiveness in that length of time is so slight that it would seem wiser to spend one's energies convincing the school board to delay the decision. Similarly, a variety of constraints beyond the evaluator's control (for example, inadequate resources, lack of administrative cooperation or support, limited time in which to collect decent evaluation data, impossible evaluation tasks, and inaccessible data essential to the evaluation) can prevent the evaluator from providing useful information. Well-intentioned but naïve clients may request "mission impossible" evaluations that yield only wasted efforts and disappointment. The evaluator needs to recognize when an evaluation is doomed to fail from the beginning. If unreasonable constraints preclude a professionally responsible evaluation, the evaluator should decline it. A bad evaluation is worse than none at all; poor evaluation data can readily mislead and lull administrators into the false security of thinking that the misinformation they have really portrays their efforts.

The Type of Evaluation Is Premature for the Stage of the Program

Programs that are in a tryout phase nearly always benefit from well-conducted formative evaluation (barring reasons listed hereafter). But one cannot be so quick to conclude that a summative evaluation would always be appropriate. Premature summative evaluations are among the most insidious misuses of evaluation, prompting concerns such as those expressed by Campbell (1984):

> Another type of mistake involved immediate evaluation, evaluation long before programs were debugged, long before those who were implementing a program believed there was anything worth imitating.
>
> When any one of them, after a year or so of debugging, feels they have something hot, a program worth others borrowing, we will worry about program evaluation in a serious sense. Our slogan would be, "Evaluate only proud programs!" (Think of the contrast with our present ideology, in which Washington planners in Congress and the executive branch design a new program, command immediate nationwide implementation, with no debugging, plus an immediate nationwide evaluation.) (pp. 35–37)

Today, the political pressures to evaluate outcomes and program impact often lead to premature summative evaluations. Programs with potentially effective models may be scuttled because of too-early summative judgments when fine-tuning those programs might result in their success. Money spent in more careful needs assessment and formative evaluation during program development and early delivery can lead to programs that are prepared for summative evaluations. Tharp and Gallimore (1979) illustrate a more effective approach to evaluation. Their approach requires a long-term commitment to using evaluation for decision making and the development of evaluation questions that are appropriate for the stage of the program and the current information needs of program developers. The process is an iterative one. Results of one study are used to make changes and refinements—with the next study examining whether these changes have succeeded.

Propriety of Evaluation Is Doubtful

Evaluations are undertaken for many reasons—some noble and some not. When the evaluator can discern that the reasons for undertaking the study are honorable and appropriate, the chances that the evaluation will be a success are enhanced. But the evaluator must also be able to recognize less noble reasons, including those that strain or violate professional principles. It would be unwise to proceed with any evaluation if its propriety is threatened by conflict of interest, jeopardy to participants in the study, or any other factors.

Propriety is one of the five areas identified by the Joint Committee on Standards for Educational Evaluation (2010) for judging the quality of an evaluation. The identification of this area, along with accuracy, feasibility, utility, and evaluation accountability, indicates the great importance professional evaluators

place on propriety. The standards outlined under propriety by the Joint Committee are designed to ensure that the evaluation will protect the rights of those involved in the evaluation, whether they be program recipients (students, clients, patients, the general public), staff, managers, or other stakeholders. An evaluation that is conducted with propriety respects the rights and dignity of those from whom data are collected and works to help organizations to address the needs of all their clients. (See Chapter 3 on ethical aspects of evaluation and Appendix A for a complete list of the *Standards* and Guiding Principles.)

Determining When an Evaluation Is Appropriate: Evaluability Assessment

As we discussed in Chapter 7, Joseph Wholey is one of the developers of decision-oriented approaches to evaluation. In that chapter, we described some of his methods. Here we will discuss more specifically how to use evaluability assessment. In the early 1970s, Wholey and his colleagues at the U.S. Department of Health, Education, and Welfare (now Health and Human Services) saw that the proliferation of program evaluation in the 1960s had not resulted in an increase in the use of program evaluation for decision making (Buchanan & Wholey, 1972). In fact, many of the potential users of evaluation were unhappy with such studies, believing that they often failed to provide useful information.

Wholey and his colleagues developed evaluability assessment as a tool to remedy this situation. They saw this as a means for facilitating communication between evaluators and stakeholders, for determining whether a program was evaluable, and for focusing the evaluation study itself.

The developers of evaluability assessment believed that many evaluations had failed because of discrepancies between "rhetoric and reality" (Nay & Kay, 1982, p. 225). As Nay and Kay point out, different levels of policymakers and program managers have different rhetorical models of the program. The models of high-level policymakers may be quite general, reflecting their role in advocating for resolution of the problem and gaining funding. The rhetorical models of managers closer to program delivery become more specific and closer to reality. Yet, even these models may fail to match reality. Many policymakers and managers may continue to cling to their rhetorical models because they perceive their particular model as necessary for public consumption. In any case, the varying rhetorical models and the gap between rhetorical models and reality make program evaluation difficult. The evaluator is unsure which program "reality" to assess.

Other common barriers to a program's being evaluable include nebulous or unrealistic goals and objectives, failure to link program activities to these goals and objectives, and managers who are unable or unwilling to make program changes on the basis of evaluation information (Horst, Nay, Scanlon, & Wholey, 1974). Other problems Wholey and others have discussed in more recent work include (1) the failure of evaluators and managers to agree on goals, objectives, and

performance criteria for measuring these objectives; (2) the inability to obtain data on program performance; and (3) problems with the particular purposes and uses of the evaluation itself (Wholey, 1983, 1986; Wholey, Hatry, & Newcomer, 2004). Wholey and his colleagues wanted to develop a way to remedy these problems.

Evaluability assessment was devised to help programs meet the four criteria deemed necessary for meaningful evaluation:

1. Program goals and priority information needs are well defined. This includes agreeing on performance criteria.
2. Program objectives are plausible. That is, there is some likelihood they can be achieved given the program's logic model or theory, the characteristics of the intended target audience, the knowledge and skills of program deliverers, and the resources provided.
3. Relevant performance data can be obtained at reasonable cost.
4. Intended users of the evaluation have agreed on how they will use the information (Wholey, 2004b, p. 34).

Evaluability assessment was first developed as a precursor to a summative evaluation; if the evaluability assessment revealed that the program did not meet the criteria, the summative evaluation would not proceed. However, the methods were not well articulated and its use declined. M. Smith (1989; 2005) then developed evaluability assessment further, using it at the U.S. Department of Agriculture in the 1980s to improve program planning. Although it is not used as frequently today, it remains an effective method for clarifying evaluation requests for many types of evaluation studies, whether used alone or in conjunction with developing a program logic model or theory or developing relations with stakeholders and potential users.

How Does One Determine Whether a Program Is Evaluable?

The major steps to determining whether a program is evaluable are:

1. Clarify the intended program model or theory.
2. Examine the program in implementation to determine whether it matches the program model and could, conceivably, achieve the program goals and objectives.
3. Explore different evaluation approaches to determine the degree to which they meet stakeholders' information needs and are feasible to implement.
4. Agree on evaluation priorities and intended uses of the study.

These steps are achieved not by the evaluator alone but in conjunction with the intended users of the study. A working group is established to clarify the program model or theory and to define their information needs and expectations for the evaluation. The role of the evaluator is to facilitate these discussions and to listen and learn about the program and the stakeholders. Wholey (1994) writes: "evaluators do not hypothesize the program design. Instead, they extract the program design . . . from relevant documentation and key actors in and around

the program" (p. 20). An evaluation can really miss an opportunity for dialogue and a critical step in understanding stakeholders' perspectives on the program if the evaluators develop their model of the program and simply assume that stakeholders agree.

What methods are used to accomplish these tasks? In addition to the facilitation of the working group, the evaluator can conduct personal interviews with stakeholders, review existing program documents (proposals, reports, brochures, and so forth), and observe the program's implementation. The interviews and program documents help the evaluator facilitate the early discussions of the working group to achieve consensus on the program model or theory.

The program model or theory should delineate the goals and objectives of the program and the principles that link them to the program actions. Frequently, a model will take the form of a flowchart, linking program actions and assumptions to the program goals and objectives. Alternative models may be developed as necessary to facilitate communication. Closure occurs when stakeholders achieve consensus on a particular model that is sufficiently detailed for the evaluator to conduct a study. (This stage is quite like developing a program logic model or theory. See Chapter 12.)

Site visits and further study of program documents (quarterly reports, resource allocations, other evaluation studies) can then help the evaluator determine (1) whether the program is being implemented according to the model, and (2) whether the implementation can feasibly achieve the desired goals. If problems occur in either of these areas, the evaluator should return to the working group and help them determine whether to revise the model to match the program reality or change the program so the program implementation corresponds to the current model. The working group can then address whether and when the evaluation should proceed. In cases where major program changes should be initiated, any outcome or summative evaluations should be postponed until program stability is achieved.

If, instead, program implementation appears to be going smoothly and the activities appear to have some chance of achieving intended outcomes, the working group can turn to examining various evaluation questions. The evaluator would also facilitate this discussion to provide guidance about what evaluation can accomplish, at what cost, and in what timeframe. By this time, the evaluator also should have learned about the various stakeholders' needs from interviews with them. Alternative evaluation plans can then be developed, specifying the questions the evaluation will answer, data to be collected, time and resources required, and potential outcomes and uses. The working group then needs to select a plan.

At any stage, the group and/or the evaluator can conclude that an evaluation is inappropriate at this time or that a quite different evaluation is required. Evaluations might be postponed for the following reasons:

- Consensus cannot be achieved among major stakeholders on the program model.
- Program actions differ greatly from the program model.

- Program actions could not feasibly achieve any stated goals or objectives of the model.
- Major stakeholders cannot achieve consensus on the direction and use of the evaluation.
- The desired evaluation plan is not feasible given data availability and resources.
- Intended uses of the evaluation are too ambiguous.

Any of these conditions might lead to the conclusion that the intended evaluation is inappropriate at that time. However, the process may have led to another type of evaluation. Specifically, the working group and/or the evaluator may conclude that, although the originally intended outcome study is inappropriate due to lack of agreement on the program model or failure in program implementation, a needs assessment or monitoring study would be useful at this time. A needs assessment study could be used to improve the program model, and a monitoring study could determine whether proposed changes in the implementation of the program occur. Thus, the process of determining when an evaluation is appropriate may result in a relatively simple "go" or "no-go," or it might result in a changed evaluation focus. In either case, through this planning effort, the evaluator has made a major step toward conducting an evaluation that makes a difference in organizational effectiveness.

Checklist of Steps for Determining When to Conduct an Evaluation

The checklist in Figure 11.1 should help the evaluator decide when to initiate an evaluation. However, when the decision is to go ahead, the evaluator may still choose to adopt some of the methods discussed to assist in focusing the evaluation.

Using an Internal or External Evaluator

In the previous section, we discussed when to conduct an evaluation. We now consider who will conduct the evaluation. The first decision might be whether to use an external or internal evaluator. When the decision to be made is summative—whether to continue, expand, or drop a program—an external evaluator (also called a third-party evaluator, independent evaluator, evaluation consultant, or evaluation contractor) may be preferable to an internal evaluator. However, as evaluation has grown as a field, we have realized that relatively few evaluations are purely summative. Most evaluations have implications for formative and summative decisions, and internal and external evaluators can present distinct differences. Note that, with the growth of performance monitoring and evaluation capacity building, internal evaluators are becoming a much more common fixture in many organizations.

FIGURE 11.1 ***Checklist for Determining When to Conduct an Evaluation***

		Check One for Each Item	
		Yes	*No*
Step 1.	*Is there a contractual requirement to evaluate?* (If yes, initiate the evaluation; if no, go to step 2.)	_____	_____
Step 2.	*Does the object of the evaluation have enough impact or importance to warrant formal evaluation?* (If yes, go to step 3; if no, formal evaluation is unnecessary, and you should discontinue further use of this checklist.)	_____	_____
Step 3.	*Is there sufficient consensus among stakeholders on the model for the program? Its goals and objectives?* (If yes, go to step 4; if no, consider a needs assessment study.)	_____	_____
Step 4.	*If the program has begun, are its actions consistent with the program model? Is achievement of goal(s) feasible?* (If yes, go to step 5; if no, consider a needs assessment or monitoring evaluation to study program modifications.)	_____	_____
Step 5.	*Is the proposed evaluation feasible given existing human and fiscal resources and data availability?* (If yes, go to step 6; if no, find more resources before proceeding or revise the scope of your plan.)	_____	_____
Step 6.	*Do the major stakeholders agree on the intended use of the evaluation?* (If yes, go to step 7; if no, discontinue or focus on those stakeholders who can use the information effectively.)	_____	_____
Step 7.	*Are the stakeholders in a position to use the information productively?* (If yes, go to step 8; if no, discontinue or focus on other stakeholders who can use the information to make decisions or take action.)	_____	_____
Step 8.	*Will the decisions of your primary stakeholders be made exclusively on other bases and be uninfluenced by the evaluation data?* (If yes, evaluation is superfluous—discontinue; if no, go to step 9.)	_____	_____
Step 9.	*Is it likely that the evaluation will provide dependable information?* (If yes, go to step 10; if no, discontinue.)	_____	_____
Step 10.	*Is the evaluation likely to meet acceptable standards of propriety?* (See Chapter 3.) (If yes, go to summary. If not, consider other means of data collection or discontinue.)	_____	_____

Summary:
Based on steps 1–10 above, should an evaluation be conducted?

Advantages of External Evaluations

The advantages of using an external agency or individual to conduct the evaluation can be summarized as follows:

1. The external evaluation is likely to be viewed as more impartial and objective because the external evaluator has more distance from the program and the people involved in its planning and implementation than the internal evaluator does.

2. The external evaluation is likely to be more credible to outside audiences, especially if the program is high profile and controversial.

3. External evaluation enables an agency to draw on evaluation expertise beyond that possessed by agency staff. Many school systems and other public and nonprofit organizations simply do not find it feasible to hire sufficient numbers of evaluation specialists to conduct the evaluations needed in the system, but they can obtain the necessary expertise through external evaluators. Moreover, external evaluators fit into more flexible staffing arrangements because there is no need for continuing financial commitment, as is the case with internal evaluators. Therefore, the particular skills of several individual external evaluators might be employed at appropriate stages, with each being paid only for the specific services needed.

4. External evaluators bring with them a new, outside perspective. Unlike the internal evaluator, they may be better able to see both the forest and the trees and may detect unwarranted assumptions that are accepted or unnoticed by insiders. Of course, external evaluators bring their own baggage based on their experiences with other similar programs, so care must be taken in hiring to learn of the external evaluator's perspectives if this area is a priority.

5. Sometimes persons associated with a program are more willing to reveal sensitive information to outsiders than they are to on-site evaluators, who they fear may inadvertently breach confidentiality because they are continually on site and in contact with others involved in the program.

6. External evaluators can feel more comfortable than internal evaluators in presenting unpopular information, advocating program changes, and working to disclose findings broadly. Specifically, because their future salaries and promotions do not depend on people in the organization, external evaluators can be as blunt and honest as the situation merits and their backbone allows. Internal evaluators can be inhibited by future concerns. (This perceived advantage of the external evaluator can, however, be overstated. External evaluators are often interested in further work with the organization and good references, if not permanent employment.)

This list is based on the ideal external evaluator, hired for a one-shot evaluation. But, in fact, many external evaluators do repeat business with an organization and, therefore, lose some of these advantages and become more like an internal evaluator.

Advantages of Internal Evaluations

Internal evaluators, too, differ broadly in their placement in the organization and their characteristics. Some internal evaluators may be employed in a separate evaluation unit, have full-time responsibilities in evaluation, and comprehensive, in-depth training in the evaluation area. (See Christie & Barela [2009] for Christie's interview with Eric Barela, a full-time, internal evaluator with a large, urban school district.) In small organizations, internal evaluations may be conducted by managers or staff whose primary responsibilities and training are in

other areas. Obviously, the internal person with more evaluation expertise is preferable to one whose expertise lies less in evaluation and more in other areas. Nevertheless, both types of internal evaluators share some advantages:

1. Internal evaluators have more knowledge of the program model and its history. This advantage can make internal evaluators quite useful in needs assessment and monitoring studies or in assessing immediate outputs or outcomes for formative purposes.

2. Internal evaluators are more familiar with the various stakeholders and their interests, concerns, and influence. This knowledge can help increase the use of the evaluation. Further, if the evaluator has formed positive relationships with management and staff, this relationship can help ease anxiety and build trust regarding evaluation.

3. Internal evaluators know the history of the organization; its clients, funders, and other stakeholders; the environment in which it operates; and the typical dynamics involved in decision making. Therefore, they can more readily and accurately identify persons who will make productive use of the study and can time and present the study to maximize its use.

4. Internal evaluators will remain with the organization after the evaluation and can continue to serve as advocates for use of its findings.

5. Because internal evaluators are already employed by the organization and oriented to it and the program, start-up times for the evaluation can be quicker than when searching for, selecting, and hiring an external evaluator who will need to take time to learn its dynamics, unless she has worked with the organization in the past.

6. Internal evaluators are a known quantity. Their strengths and weaknesses are known to the organization and can be analyzed in reference to the project under consideration (Love, 1991; Sonnichsen, 1999).

The definition of an internal evaluator becomes less clear in larger organizations and governmental units. Evaluators from the U.S. Government Accountability Office (GAO) would probably be considered external evaluators when they evaluate an executive branch program in response to a congressional request, even though they are federal employees evaluating a federal program. Is an employee of a state evaluation unit or state auditing office considered an internal evaluator because he or she is employed by the state? Not likely, although some citizens might be concerned with the independence of the study regarding a controversial program or policy. What if the evaluator is part of an evaluation unit within the state organization that manages the program to be evaluated? In that case, the evaluator would more likely be considered an internal evaluator, especially if the organization were a small one.

The prototypical internal evaluator is an employee of a small organization who works daily with program planners and providers. Moderate-size nonprofit organizations and many units of local government are examples of organizations that would include such internal evaluators. Conversely, the prototypical external

evaluator is an independent consultant or an employee of an organization whose function is to conduct evaluations by contract. Many evaluators lie somewhere between these two extremes. Nevertheless, the contrasts between internal and external evaluators, like the distinctions between summative and formative evaluations, help us to examine the strengths and weaknesses of the various evaluators we could select to conduct the study. We also can use the continuum from internal to external evaluator to ameliorate some of our concerns. For example, the concern regarding impartiality or bias of the internal evaluator may be partially remedied by selecting an internal evaluator who is relatively distant, on an organizational chart, from the program. Note, however, that this distance, while improving impartiality, diminishes the internal evaluator's typical advantage of knowing the program and its stakeholders.

Sonnichsen (1999), formerly the Director of the FBI's Office of Planning and Evaluation, describes how internal evaluation units can be established and organized to maximize impact. He sees internal evaluators as having the potential to "build an organizational tradition of systematic, critical review and reflection on organizational issues and problems with positive consequences in terms of improved performance" (Sonnichsen, 1999, p. 2). He proposes five preconditions for internal evaluators to have high impact on an organization: supportive top management, availability of competent evaluators, an organizational culture of internal review, reliable data systems, and unlimited access by evaluators to the organization's data and personnel. His successful internal evaluator is very much oriented to improving the decision-making process within organizations. (See also Love [1991] for more on internal evaluation.)

Another recent model for internal evaluators is one focused on evaluation capacity building (ECB). In this role, the evaluator is less concerned with conducting individual evaluation studies and more concerned with creating and sustaining an environment conducive to evaluation studies and their use within the organization. (See discussion of ECB in Chapter 9.)

Advantages of Combining Internal and External Evaluation

Internal and external evaluation are far too often viewed as mutually exclusive. They need not be. Combining the two approaches can compensate for several of the disadvantages of each that were mentioned previously. The external evaluator's lack of familiarity with the program and its stakeholders is less of a problem if she works in tandem with an internal evaluator who can provide the necessary contextual information. Travel costs can be greatly reduced by having the internal evaluator collect the bulk of the necessary data and actively communicate evaluation plans and results to significant internal audiences. Finally, after the external evaluator is gone, the internal evaluator will remain as an advocate for the use of the evaluation.

The external evaluator can then be used to increase impartiality and credibility as well as to provide specialized knowledge and skills that are not routinely needed in-house. The external evaluator can assist with key tasks in which bias

might inadvertently occur, such as designing the evaluation, selecting or developing instruments, drawing conclusions from data, and the like. The external evaluator can interpret and present sensitive results to stakeholders.

External evaluators can also be used to "audit" internal evaluation studies to certify that they are methodologically sound and unbiased (Chen, 1994; Sonnichsen, 1999). Such partnerships incorporate the advantages of external evaluation without requiring that the entire evaluation be conducted externally. Further, through the resulting teamwork, internal evaluators can learn new evaluation methods to use in the future.

Checklist of Steps for Determining Whether to Use an External Evaluator

Figure 11.2 is proposed as a checklist for deciding whether or not to use an external agency or individual to conduct the evaluation.

FIGURE 11.2 ***Checklist for Determining Whether to Use an External Evaluator***

		Check One for Each Item	
		Yes	*No*
Step 1.	*Is there a contractual requirement that the evaluation be conducted by an external evaluator?* (If yes, initiate the search for an external evaluator; if no, go to step 2.)	_____	_____
Step 2.	*Are financial resources available to support the use of an external evaluator?* (If yes, proceed to step 3; if no, discontinue use of this checklist and conduct the evaluation internally.)	_____	_____
Step 3.	*Does the evaluation require specialized knowledge and skills beyond the expertise of internal evaluators who are available to do the evaluation tasks?* (If yes, initiate the search for an external evaluator; if no, go to step 4.)	_____	_____
Step 4.	*Is the evaluation concerned with measuring major or highly politicized goals for summative purposes?* (If yes, initiate the search for an external evaluator; if no, go to step 5.)	_____	_____
Step 5.	*Is an outside perspective of particular importance to the study?* (If yes, initiate the search for an external evaluator; if no, go to the summary.)	_____	_____

Summary:
Based on steps 1–5 above, should this evaluation be conducted by an external evaluator?

Hiring an Evaluator

Hiring an evaluator, whether as a permanent internal employee or as an external consultant, is neither simple nor trivial. There is no better way to guarantee a bad evaluation than to turn it over to someone who is inept. Relationships with stakeholders can be irreparably harmed by an insensitive or unresponsive evaluator. Misleading or incorrect information is easy to generate and disseminate but difficult to eradicate. Therefore, great care should be exercised in hiring evaluators. Before summarizing some criteria that have been suggested for hiring evaluators, it is necessary to consider briefly what competent evaluators must be able to do.

Competencies Needed by Evaluators

There have been several conceptual and/or empirical efforts to identify the tasks required of evaluators and the more specific competencies (knowledge, skills, and sensitivities) required to perform those tasks well (e.g., Covert, 1992; King et al., 2001; Mertens, 1994; Stevahn, King, Ghere, & Minnema, 2005; Worthen, 1975).

The overlap among various lists is reassuringly high. We would be concerned if there were substantial disagreement among professional evaluators regarding critical competencies, but that is not the case. In fact, in their study of professional evaluators' consensus on competencies, King et al. (2001) found substantial agreement.

The few areas where the lists of competencies do not overlap result, we believe, from: (1) different publication dates (new issues and needs in evaluation are often being discovered); (2) differences in level of detail; and (3) differences in the evaluation philosophy or setting of the authors. For example, King et al. (2001) found that the areas in which consensus did not emerge often reflected differences in the context and role of the different evaluators taking part in their study and, thus, may reflect different types of evaluation practice. For example, conflict-resolution skills were viewed as more important by evaluators who work closely with stakeholders.

Most recently, King et al. (2001) and Stevahn et al. (2005) have used research, presentations, discussions in the professional community, and reflection to develop a set of competencies for program evaluators that could be used for planning the education and training of evaluators, for screening and hiring evaluators, and for evaluators' own reflective practice. Their initial process used Multi-Attribute Consensus Reaching (MACR) procedures to develop a list of competencies in collaboration with 31 professional evaluators who worked in different settings (King et al., 2001). Subsequently, they presented these competencies at professional conferences and sought and received input from more than 100 professional evaluators. To communicate to the readers who are new to evaluation what they should aspire to learn, we present the list in Figure 11.3.

FIGURE 11.3 ***Taxonomy of Essential Competencies for Program Evaluators***

1. Professional practice

- 1.1 Applies professional program evaluation standards
- 1.2 Acts ethically and strives for integrity and honesty in conducting evaluations
- 1.3 Conveys personal evaluation approaches and skills to potential clients
- 1.4 Respects clients, respondents, program participants, and other stakeholders
- 1.5 Considers the general and public welfare in evaluation practice
- 1.6 Contributes to the knowledge base of evaluation

2. Systematic inquiry

- 2.1 Understands the knowledge base of evaluation (terms, concepts, theories, and assumptions)
- 2.2 Knowledgeable about quantitative methods
- 2.3 Knowledgeable about qualitative methods
- 2.4 Knowledgeable about mixed methods
- 2.5 Conducts literature reviews
- 2.6 Specifies program theory
- 2.7 Frames evaluation questions
- 2.8 Develops evaluation designs
- 2.9 Identifies data sources
- 2.10 Collects data
- 2.11 Assesses validity of data
- 2.12 Assesses reliability of data
- 2.13 Analyzes data
- 2.14 Interprets data
- 2.15 Makes judgments
- 2.16 Develops recommendations
- 2.17 Provides rationales for decisions throughout the evaluation
- 2.18 Reports evaluation procedures and results
- 2.19 Notes the strengths and limitations of the evaluation
- 2.20 Conducts metaevaluations

3. Situational analysis

- 3.1 Describes the program
- 3.2 Determines program evaluability
- 3.3 Identifies the interests of relevant stakeholders
- 3.4 Serves the information needs of intended users
- 3.5 Addresses conflicts
- 3.6 Examines the organizational context of the evaluation
- 3.7 Analyzes the political considerations relevant to the evaluation

FIGURE 11.3 ***Continued***

3.8 Attends to issues of evaluation use
3.9 Attends to issues of organizational change
3.10 Respects the uniqueness of the evaluation site and client
3.11 Remains open to input from others
3.12 Modifies the study as needed

4. Project management

4.1 Responds to requests for proposals
4.2 Negotiates with clients before the evaluation begins
4.3 Writes formal agreements
4.4 Communicates with clients throughout the evaluation process
4.5 Budgets an evaluation
4.6 Justifies cost given information needs
4.7 Identifies needed resources for evaluation, such as information, expertise, personnel, instruments
4.8 Uses appropriate technology
4.9 Supervises others involved in conducting the evaluation
4.10 Trains others involved in conducting the evaluation
4.11 Conducts the evaluation in a nondisruptive manner
4.12 Presents work in a timely manner

5. Reflective practice

5.1 Aware of self as an evaluator (knowledge, skills, dispositions)
5.2 Reflects on personal evaluation practice (competencies and areas for growth)
5.3 Pursues professional development in evaluation
5.4 Pursues professional development in relevant content areas
5.5 Builds professional relationships to enhance evaluation practice

6. Interpersonal competence

6.1 Uses written communication skills
6.2 Uses verbal/listening communication skills
6.3 Uses negotiation skills
6.4 Uses conflict-resolution skills
6.5 Facilitates constructive interpersonal interaction (teamwork, group facilitation, processing)
6.6 Demonstrates cross-cultural competence

Source: Material is drawn from "Establishing Essential Competencies for Program Evaluators" (pp 49–51) by L. Stevahn, J. A. King, G. Ghere, and J. Minnema, 2005. *American Journal of Evaluation, 26*(1), 43–59. Reprinted with permission.

Note that the list contains many of the factors that we discuss in this chapter, such as negotiating with clients before the evaluation begins, determining program evaluability and theory, examining the organizational context and political considerations of the evaluation, and identifying the interests of relevant stakeholders.

We will now turn to the subject of hiring an evaluator. The competencies, however, will prove useful in reminding the manager or person doing the hiring that methodological skills are only one element of being a successful evaluator. Interpersonal skills, communication skills, and management skills must also be examined. Many successful methodologists, for example, struggle in communicating complex results to audiences unfamiliar with research methods or lack the interpersonal skills to work successfully—listening, understanding, and resolving conflicts—among a variety of stakeholders.

Possible Approaches to Hiring an Evaluator

What are the means by which an agency can determine whether an evaluator has these competencies? As in any personnel process, selection methods should be matched to the knowledge and skills needed for the job. A résumé and/or past evaluation reports can be useful in judging whether the candidate has the necessary methodological expertise and writing skills. An interview with the candidate—if possible, conducted by representatives of different stakeholders—can be used to assess the candidate's oral communication skills and ability to work with different audiences. An interview can be particularly successful in determining an evaluator's ability to explain complex issues clearly (in describing previous work) and to listen and learn. The candidate's questions and comments during an interview can be judged in terms of the applicant's interest in the program and the evaluation, sensitivity to different stakeholders, and overall oral communication skills. Finally, talking with others who have used the evaluator can be invaluable in discovering more about the candidate's skills in managing an evaluation responsibly and ethically. Such references can also provide useful information about the personal style and professional orientation of the evaluator, as can samples of any reports produced for past clients who are willing to share them.

Checklist of Questions to Consider in Selecting an Evaluator

Figure 11.4 is designed as a checklist of criteria to consider in selecting an evaluator. It is intended to build on the competencies listed in Figure 11.3 with the person completing the list basing his or her conclusions on the competencies relevant to that question.

FIGURE 11.4 ***Checklist of Questions to Consider in Selecting an Evaluator***

		Evaluator Qualifications (Check One for Each Item)		
		Yes	*No*	*?*
Question 1.	*Does the evaluator have the ability to use the methodologies and techniques that may be required in the study?* (Consider education and training, past experience, and philosophical orientation.)			
Question 2.	*Does the evaluator have the ability to help articulate the appropriate focus for the study?* (Consider communication skills, ability to work with stakeholder groups, content specialization.)			
Question 3.	*Does the evaluator have the management skills to carry out the study?* (Consider education and training, past experience.)			
Question 4.	*Will the evaluator maintain appropriate ethical standards?* (Consider education, training; talk with references.)			
Question 5.	*Will the evaluator be interested in and able to communicate results to desired stakeholders in such a way that the results will be used?* (Examine previous evaluation documents; talk with references.)			

Summary:
Based on questions 1–5 above, to what extent is the potential evaluator qualified and acceptable to conduct the evaluation?

How Different Evaluation Approaches Clarify the Evaluation Request and Responsibilities

How would the proponents of the different models reviewed in Chapters 5 through 8 approach the clarification of the evaluation request? Most proponents of the models described in those chapters would not object to the methods discussed in this chapter. All except Scriven's goal-free evaluation would include at least some interviews with stakeholders and reviews of existing documents during the planning stage to clarify the request.

An evaluator subscribing to a program-oriented model would focus primarily on the specification of objectives or a program theory or logic model during this stage. Contemporary evaluators would be more likely to be theory based. As

such, they would spend more time during the planning stage working with the client and stakeholders to learn or develop their normative theories for the program and would review relevant research literature to identify theories and research findings that might be explored or used in the evaluation.

In contrast, the decision-oriented evaluator would focus more on the decisions to be made and the information needs of the managers who would make those decisions. If the decisions concerned program theories or objectives, the evaluator would focus on those; however, if the decisions concerned other issues, the evaluator would readily adapt. The decision-oriented evaluator would be interested in learning who was interested in the evaluation and who had used evaluations in the past, so as to identify potential primary users. He or she would also consider the stage of the program and the likely information needs. In contrast to the decision-oriented approach, the consumer-oriented evaluator would clarify the request by conducting a functional analysis of the product and identifying the criteria to be used to judge its merit or worth. Similarly, the expertise-oriented evaluator might limit the evaluation by relying on established areas of concern in the discipline of the program (e.g., medical, education, or environmental standards). The clarification might have little focus on the needs of actual stakeholders, as the evaluator would assume that she was hired to define the criteria based on her personal expertise in the discipline of the program.

The participant-oriented evaluator might either involve many stakeholders—more than adherents of other models would—in the clarification of the evaluation request or focus on a particular group of stakeholders (typically, program managers and staff) to learn their perspectives and expectations regarding the program and the evaluation and to begin to engage them in the evaluation itself. If the participant-oriented evaluation would involve many different stakeholders, the evaluator should begin to examine differences in values, perspectives, and power of these groups. Meetings of diverse working groups, as proposed by Wholey in evaluability assessment, might be more acrimonious, and might make it more difficult to achieve consensus as to the purpose of the evaluation and the likely procedures to be used. Further, if the gap between managers or policymakers and other stakeholders were too great, an external evaluator (because he would be less likely to be perceived as a tool of management) with strong conflict-resolution skills might be more appropriate. However, such working groups could succeed in identifying issues that might not have been raised in managerial meetings and could further communication between managers and other stakeholders. As an alternative, the participant-oriented evaluator might clarify the evaluation request through interviews with different stakeholder groups, reviews of program documents, and observations of the program, without trying to achieve consensus through a working group of all stakeholders. The participant-oriented evaluator who is of a constructivist bent would be looking for different views, or multiple realities, not necessarily consensus.

All approaches can be implemented by either internal or external evaluators; however, internal evaluators are more likely to be decision oriented. Their primary purpose is to assist managers in decision making. Because of their ongoing presence, their focus is more on organizational improvement than on any

individual program. They would be less likely to take a consumer-oriented or expertise-oriented approach because their primary stakeholder is unlikely to be a consumer. They have been hired for evaluation expertise, not expertise in the content of the particular program. External evaluators, in contrast, are typically hired to evaluate a specific program or policy. They may be hired by consumers, managers, or any group of stakeholders for their evaluation expertise or their expertise in the content of the particular program they are to evaluate. Within that context, elements of theory-based and participatory approaches may be adapted by either internal or external evaluators, given the particular circumstances of the evaluation and needs of the stakeholders.

Major Concepts and Theories

1. Evaluations can be used by the sponsor, client, audience, or stakeholders. Each group has its own needs, concerns, and informational requirements for the evaluation. The evaluator should identify each group and, as appropriate, incorporate their concerns in the evaluation plan.

2. Determining and understanding the purpose for an evaluation is probably the most important activity to be completed before the evaluation begins.

3. Evaluation can serve many uses, including direct information use, educating users about alternative ways to make decisions, stimulating dialogue among stakeholders, and raising awareness of program issues or stakeholder views.

4. It may be inappropriate to conduct an evaluation if the client is using it to avoid responsibility, for public relations purposes, or to postpone making a decision; resources are inadequate; trivial or invalid information will be produced; the evaluation could lead to unethical practices; or audiences will be misled.

5. Evaluability assessment can be used to determine if it will be effective to proceed with an evaluation. This includes working with program managers to determine if goals and program models or theories are clearly articulated and feasible and if identified audiences will use the information.

6. Internal or external evaluators can conduct evaluations. Internal evaluators have the advantage of knowing the organization, its history, and its decision-making style and will be around to encourage subsequent use. External evaluators can bring greater perceived objectivity and specialized skills for a particular project.

Discussion Questions

1. Why is it important to clarify the evaluation requests? What do we mean by that?

2. How might the typical information needs of sponsors, clients, audiences, and other groups of stakeholders differ?

3. How do you think evaluability assessment might help the evaluator? The users?

4. Which competencies in Figure 11.3 are new or surprising to you? Which do you think are most important? Would you add any competencies to the list?
5. Under what circumstances would you prefer to use an internal evaluator? Name a program or issue that you would prefer an internal evaluator to address. Do the same with an external evaluator. What concerns would you have about each?

Application Exercises

1. What questions might you want to ask if you were being considered to perform an evaluation?
2. Consider a program you know. Does it meet Wholey's criteria for evaluability? If not, what changes should occur? Are there any steps you could take as an evaluator to help achieve these changes?
3. Considering the program you identified in Exercise 2 and a probable need for information, would an internal or external evaluator be preferable for this evaluation? Justify your choice.
4. What competencies would be most important to evaluate the program you are considering? How would you go about hiring a person to conduct this evaluation (internal or external)?

Relevant Evaluation Standards

We consider the following evaluation standards, which are listed in their entirety in Appendix A, to be relevant to this chapter's content:

U1—Evaluator Credibility
U2—Attention to Stakeholders
U3—Negotiated Purposes
U4—Explicit Values
F3—Contextual Viability
P1—Responsive and Inclusive Orientation
P6—Conflicts of Interests
E1—Evaluation Documentation
A2—Context Analysis
A3—Described Purposes and Procedures

Case Studies

For this chapter, we recommend reading three interviews that illustrate negotiating different roles of internal and external evaluators. *Evaluation in Action*, Chapter 3, is an interview with Len Bickman, an external evaluator, and Chapter 5 presents an interview with David Fetterman in a semi-internal role. The third interview is with an internal evaluator, Eric Barela, and is only available in *American Journal of Evaluation,* as it was conducted after the publication of *Evaluation in Action.*

In Chapter 3, Len Bickman discusses his disagreements with one of the sponsors of the evaluation and his insistence on ownership of some critical evaluation decisions in order to maintain the validity of the study. The journal source is Fitzpatrick, J. L., & Bickman, L. (2002). Evaluation of the Ft. Bragg and Stark County systems of care for children and adolescents: A dialogue with Len Bickman. *American Journal of Evaluation*, *23*(1), 67–80.

In Chapter 5, David Fetterman frankly discusses some of the problems he encounters with faculty colleagues in evaluating a program in his own School of Education, although he is not directly connected with the program. The journal source is Fitzpatrick, J. L., & Fetterman, D. (2000). The evaluation of the Stanford Teacher Education Program (STEP): A dialogue with David Fetterman. *American Journal of Evaluation, 21*(2), 240–259.

The third interview we recommend is with Eric Barela, an internal evaluator in a large, urban school district. An internal evaluator like Fetterman, his interview indicates how a formal internal office may have protections that differ from those for an individual internal evaluator undertaking a project. The journal source is Christie, C. A., & Barela, E. (2009). Internal evaluation in a large urban school district: A Title I best practices study. *American Journal of Evaluation*, *29*(4), 531–546.

Suggested Readings

Joint Committee on Standards for Educational Evaluation. (2010). *The program evaluation standards* (3rd ed.). Thousand Oaks, CA: Sage.

Smith, M. F. (1989). *Evaluability assessment: A practical approach*. Boston: Kluwer Academic.

Smith, N. (1998). Professional reasons for declining an evaluation contract. *American Journal of Evaluation*, *19*, 177–190.

Sonnichsen, R. C. (1999). *High impact internal evaluation*. Thousand Oaks, CA: Sage.

Stevahn, L., King, J. A., Ghere, G., & Minnema, J. (2005). Establishing essential competencies for program evaluators. *American Journal of Evaluation*, *26*(1), 43–59.

Wholey, J. S. (2004). Evaluability assessment. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (2nd ed.). San Francisco: Jossey-Bass.

12

Setting Boundaries and Analyzing the Evaluation Context

Orienting Questions

1. Who are the potential stakeholders and audiences for an evaluation? When and how should the stakeholders be involved in the evaluation?
2. Why is it important to describe the object of the evaluation?
3. What function do logic models and program theories serve? What steps does the evaluator take to develop a program theory?
4. How can knowledge of potential resources for the evaluation help in the planning at this stage?
5. What should the evaluator consider in analyzing the political context in which an evaluation will occur? What impact would political considerations have on the conduct of the study?

In the preceding chapter, we dealt with determining whether to conduct an evaluation, deciding whether to use an internal or external evaluator, and considering whom to hire. In this chapter, we turn our attention to four other important considerations: identifying evaluation stakeholders and audiences, setting boundaries on whatever is evaluated, learning more about the program through developing logic models and a program theory, and analyzing available resources and the political context.

Identifying Stakeholders and Intended Audiences for an Evaluation

During the planning stage of an evaluation, it is essential that the evaluator identify all the various stakeholders and audiences for the evaluation. Involving the stakeholders during the planning stage helps ensure that the evaluation addresses appropriate concerns, and it assists the evaluator in identifying potential users. Further, involving stakeholders at an early stage can help reduce their anxieties about the evaluation and allows the evaluator to learn how different groups perceive the program. Recognition of audiences beyond the immediate stakeholders can also help the evaluator to consider future dissemination of results. In this section, we discuss the identification and involvement of these groups.

Identifying Stakeholders to Be Involved in the Evaluation and Future Audiences

At the outset, the evaluator must realize that the sponsor and client—the people the evaluator usually meets first—generally represent a primary audience for the study, but there are almost always other important stakeholders and audiences to consider. Indeed, the evaluation's sponsor often provides the funding for the study to provide information for other stakeholders, such as the managers or staff of the program being evaluated in formative evaluations.

The evaluator should identify and communicate with each stakeholder group or its representative to learn that group's perceptions and concerns about the program and the evaluation. Obviously, because some stakeholders have more concerns about the program or more immediate uses for the evaluation, some weighting of their input will be necessary. In almost all evaluations, however, the final evaluation plan will include questions that address the information needs of several different stakeholder groups. So, how does one identify all the legitimate audiences?

Working with the evaluation client and/or sponsor, the evaluator must strike a reasonable balance in deciding whether to define stakeholders broadly or narrowly. Few evaluations hold sufficient interest to warrant news releases in the *Wall Street Journal* or the *London Times*, but the more frequent mistake is settling on too narrow a range. Program managers and their staff, who represent those working in the trenches, are usually involved to some extent in most evaluations. Community members and representatives of other influence groups are increasingly numbered among the evaluation's stakeholders or audiences. There is still a regrettable tendency, however, to respond to the squeaking wheel—targeting evaluation studies to those who are vociferous, strident, or powerful. What about people without school-age children who tend to think schools are not working today? What about high school students and their parents? In their model of deliberative democracy in evaluation, House and Howe (1999) argue that the evaluator has a responsibility to "use procedures that incorporate the views of insiders and outsiders [and] give voice to the marginal and excluded" (p. xix). While they acknowledge their approach as an ideal, their point is that evaluators can play a powerful role in

bringing about a "democratic dialogue" or facilitating a rational discussion on the purposes of the evaluation among groups that often do not exchange views.

Increasing the number and diversity of stakeholders can add to the complexity and cost of the evaluation. However, for political, practical, and ethical reasons, the evaluator can ill afford to ignore certain constituents. Therefore, the question of who the audiences are and how they are to be involved or served is a crucial one.

Greene (2005) has identified four groups of stakeholders:

(a) people who have decision authority over the program, including other policy makers, funders, and advisory boards;
(b) people who have direct responsibility for the program, including program developers, administrators in the organization implementing the program, program managers, and direct service staff;
(c) people who are the intended beneficiaries of the program, their families, and their communities; and
(d) people disadvantaged by the program, as in lost funding opportunities (2005, pp. 397–398).

Scriven (2007) adds others to this list, including political supporters or opponents, elected officials, community leaders, and the public in general. We have adapted Greene's categories and Scriven's suggestions to develop the checklist shown in Figure 12.1.

It is doubtful that any one evaluation would address all the stakeholders and the additional audiences listed in Figure 12.1, but the interests of each of these groups should be considered during the planning stage. The evaluator can review the categories with the client to identify stakeholders or groups who might have been forgotten but may provide a useful new perspective or who might be able to use or disseminate the evaluation in different ways or to different audiences. Once the stakeholders to be included have been determined, the client and others can recommend representatives of each group for the evaluator to meet with and discuss their perceptions of both the program to be evaluated and the evaluation itself. What does each group perceive as the purpose of the program? How well do they think it works? What concerns do they have about it? What would they like to know about it? What have they heard about the evaluation? Do they agree with its intent? What do they hope to learn from the evaluation? What concerns do they have about it? The evaluator should attempt to meet with stakeholders with diverse opinions of the program, not only to include a broad range of stakeholders, but also to gain a more complete picture of the program.

After meeting with representatives of all stakeholders who might be included, the evaluator can then make decisions, possibly with the client, about the importance and role of each group in the evaluation. Most evaluations can make effective use of stakeholders as part of an ongoing advisory group to give feedback or to assist with many of the evaluation tasks. Some might become involved in data collection and interpretation of results; others might be briefed on a more intermittent basis. Still other stakeholder groups may have little or no interest in the study, given its focus.

FIGURE 12.1 ***Checklist of Potential Stakeholders and Audiences***

Evaluation Audience Checklist

Entity to Be Evaluated *(Check all appropriate boxes)*

Individuals, Groups, or Agencies Needing the Evaluation's Findings	*To Make Policy*	*To Make Operational Decisions*	*To Provide Input to Evaluation*	*To React*	*For Interest Only*
Developer of the program					
Funder of the program					
Person/agency who identified the local need					
Boards/agencies who approved delivery of the program at local level					
Local funder					
Other providers of resources (facilities, supplies, in-kind contributions)					
Top managers of agencies delivering the program					
Program managers					
Program deliverers					
Sponsor of the evaluation					
Direct clients of the program					
Indirect beneficiaries of the program (parents, children, spouses, employers)					
Potential adopters of the program					
Agencies who manage other programs for this client group					
Groups excluded from the program					
Groups perceiving negative side effects of the program or the evaluation					
Groups losing power as a result of use of the program					
Groups suffering from lost opportunities as a result of the program					
Public/community members					
Others					

The checklist in Figure 12.1 is intended to help evaluators and clients think in broad terms about the purpose stakeholders might serve if involved in the study. Once the appropriate groups or individuals have been identified, the list should be reviewed periodically as the evaluation progresses, because stakeholders can

change as a result of personnel changes or changes in the political context of the study or the program.

As data collection plans are developed and data are collected and analyzed, it is important to consider what information each stakeholder group needs and will use and the interests of audiences beyond the stakeholders. All groups are not interested in the same information. Program deliverers and managers will be interested in more detail than will the general public or policymakers. Differing interests and needs generally require that evaluation reports be tailored for specific audiences. We will discuss this further in Chapter 17.

Importance of Identifying and Involving Various Stakeholders

The viewpoints of various stakeholders provide focus and direction for the study. Unless evaluators direct the evaluation clearly at the intended users from the outset, results are likely to have little impact. Discussing who will use evaluation results, and how, is essential to clarify the purpose of the study.

As noted in Chapter 11, most evaluators have at some time been misled (perhaps inadvertently) into undertaking an evaluation, only to find at some point that its underlying purpose was quite different from what they had supposed. Such misunderstanding is much more likely if an evaluator talks to only one stakeholder. Dialogue with multiple stakeholders also clarifies the reasons behind an evaluation.

Describing What Is to Be Evaluated: Setting the Boundaries

Setting boundaries is a fundamental step in gaining a clear sense of what an evaluation is all about. Developing a description of the program is essential in helping to establish those boundaries. Poor or incomplete descriptions can lead to faulty judgments—sometimes about entities that never really existed. For example, the concept of team teaching fared poorly in several evaluations, resulting in a general impression that team teaching is ineffective. Closer inspection showed that what was often labeled as "team teaching" provided no real opportunities for staff members to plan or work together in direct instruction. Obviously, better descriptions of the program theory and program in action would have precluded these misinterpretations. One can only evaluate adequately what one can describe accurately.

The importance of good description increases in proportion to the complexity and scope of what is evaluated. Evaluators are frequently asked to help evaluate entities as vague as "our Parks and Recreation program." Does that include all programs across all seasons, only the summer recreational programs, or only

swimming programs? Would such an evaluation focus on training of part-time summer employees, public use of parks, maintenance of parks, or all of the above? Would it determine whether the goals of Parks and Recreation meet the needs of the community, whether the program managers are correctly and effectively implementing the policies determined by elected officials, or both? Answering such questions establishes boundaries that help the evaluation make sense.

A program description explains the critical elements of a program. Such a description typically includes goals and objectives, critical program components and activities, descriptions of the target audience, and a logic model, program theory, or both. The program description may also include characteristics or expectations of the staff delivering the program, administrative arrangements, the physical setting, and other contextual and resource factors. Many descriptions provide information about extensive, delineating critical factors in the history of the program and reasons for choices made at various stages. Others are briefer, but still convey a picture of the essence of the current program. The critical factor in a program description is that it is sufficiently detailed to provide the evaluator with an understanding of why the program is supposed to achieve its desired impacts and to serve as a foundation for identifying evaluation questions. However, some descriptions are so microscopically detailed and cluttered with trivia that it becomes difficult for the evaluator to identify critical elements and linkages between program activities and outcomes. An accurate final description agreed upon by all stakeholders provides a common understanding of the program for all the parties involved, permitting the evaluation to proceed with some consensus concerning the entity to be examined.

Factors to Consider in Characterizing the Object of the Evaluation

The evaluator can demarcate the object of the evaluation and the study itself by answering a series of questions:

- What problem was the program designed to correct? What need does the program exist to serve? Why was the program initiated? What is the program's history? What are its goals? Whom is it intended to serve?
- What does the program consist of? What are its major components and activities, its basic structure and administrative/managerial design? How does it function? What research exists to link the activities of the program and characteristics of the clients with the desired outcomes?
- What is the program's setting and context (geographical, demographic, political, level of generality)?
- Who participates in the program (direct and indirect participants, program deliverers, managers and administrators, policymakers)? Who are other stakeholders?
- When and under what conditions is the program to be implemented? How much time is it intended to take? How frequently is it to be used?

- Are there unique contextual events or circumstances (e.g., contract negotiations, budgetary decisions, changes in administration, elections) that could affect the program in ways that might distort the evaluation?
- What resources (human, materials, time) are consumed in using the program?
- Has the program been evaluated previously? If so, what were the findings? How were they used?
- What critical decisions are key stakeholders facing in regard to the program? What is the timeframe for these decisions?

The evaluator should also seek to clarify what is not included in the program to be evaluated.

Using Program Theory and Logic Models to Describe the Program

We discussed theory-driven evaluation, its history, principles, and foundations and some issues concerning its implementation in Chapter 6. Here, our focus is on describing how to use logic models or program theory to understand the rationale for the program and its key components and to serve as an aid to guide the evaluation.

Since the early 1990s, the use of program theories and logic models has increased enormously (Rogers, 2007). Some of that increase was stimulated by the work of Carol Weiss with the Aspen Institute (1995) and others (Weiss, 1997) prompting nonprofit organizations and funding sources to require program theory as a foundation to evaluation and program planning. Models of theory-based evaluation were published and received significant discussion (Bickman, 1987, 1990; Chen, 1990). Mark Lipsey's (1993) article on "small theories" and their use in evaluation methodology also had a marked influence. At the same time, the U.S. federal government, through the Government Performance and Results Act (GPRA), began requiring federal agencies and those receiving federal funding to report results. The nature of this work stimulated the growth of logic models. As Taylor-Powell and Boyd (2008) write, training people in their agency on the new GPRA input-output-outcome model and terminology was a powerful impetus for their organization and others to start using logic models for planning and evaluation.

Of course, many of these efforts were not voluntary and were preceded by efforts in other countries. State and federal governments in Australia and many international development agencies had begun requiring logic models, or forms thereof, in the 1970s (Rogers, 2007). GPRA, the United Way, and many states in the United States began mandating logic models in the mid-1990s. Today, logic models or program theories are a common aspect of program development in many organizations, but certainly not all. Of course, if a program has a logic model or program theory, the evaluator can learn about it, read documents concerning it, and talk with others to learn more about the program and determine whether it should be revised to meet the planning needs for the evaluation. If a logic model or program theory does not exist, the evaluator may choose to use these methods to describe the program and to identify areas of uncertainty or information needs for the evaluation.

Many note that the terms "logic model" and "program theory" are used interchangeably and, in practice, can be identical (McLaughlin & Jordan, 2004; Weiss, 1997). But, as we noted in Chapter 6, logic models are more likely to reflect a chain of program activities, whereas program theory should help explain the reasons why the program should work. Both can play a part in helping the evaluator to describe the program and to begin to set boundaries for the evaluation.

Evaluators, however, should be clear as to the purposes of their model. Weiss (1997) delineates between two types of program theory: implementation theory and programmatic theory. Implementation theory, like a logic model, is characterized by a description or a flow chart of program inputs, activities, outputs, and outcomes. Different words or different types of outputs and outcomes may be labeled as such, for example, short-term outcomes, intermediate-term outcomes, long-term outcomes. Nevertheless, the theory is one of a sequence and may, or may not, provide any clue as to the reasons why the outcomes should be achieved. Implementation theory, of course, could be useful if the evaluation were to focus on describing the program process, but even then, the evaluator must select the important processes to describe and, typically, those are ones closely linked to outcomes. As Weiss emphasizes, implementation theory is not useful for understanding the causal mechanisms of the program, and at this stage, the evaluator does want to understand the causal mechanisms of the program. Why? Because the evaluator is still learning about the context, working to understand the program and its rationale. Such an understanding will later help the evaluator in many critical decisions—what is important to measure, when to measure it, and how to measure it (Leviton, 2001; Lipsey, 1993; Weiss, 1997).

Building a Logic Model. Let us provide some advice and resources here regarding both logic models and program theory. The University of Wisconsin Extension Program Development and Evaluation Office has created a well-known web site that includes training modules and examples of logic models. (See http://www.uwex.edu/ces/pdande/evaluation/evallogicmodel.html.) In Figures 12.2 and 12.3, we present two examples of their work. Figure 12.2 is a logic model for evaluation capacity building in an organization (Taylor-Powell & Boyd, 2008). We use this example because, having read about evaluation capacity building or ECB in Chapter 9, all readers are familiar with its concepts and tenets. The reader can see how Taylor-Powell and Boyd's logic model for an ECB Theory of Change is useful in illustrating the process in a succinct, linear manner. It highlights how ECB leads first to changes in individuals, then to changes in the program being evaluated and the team working with the program, and then to the organization as a whole and, ultimately, to society. Further, the logic model presents indicators of those changes that might be selected by an evaluator to measure progress or success in ECB.

Figure 12.3 presents a more complex logic model for reducing and preventing tobacco use by teens. This logic model could be used by an evaluator to promote dialogue among stakeholders about boundaries: In what areas do they have the most information needs? Where do they have the strongest concerns? But the theory

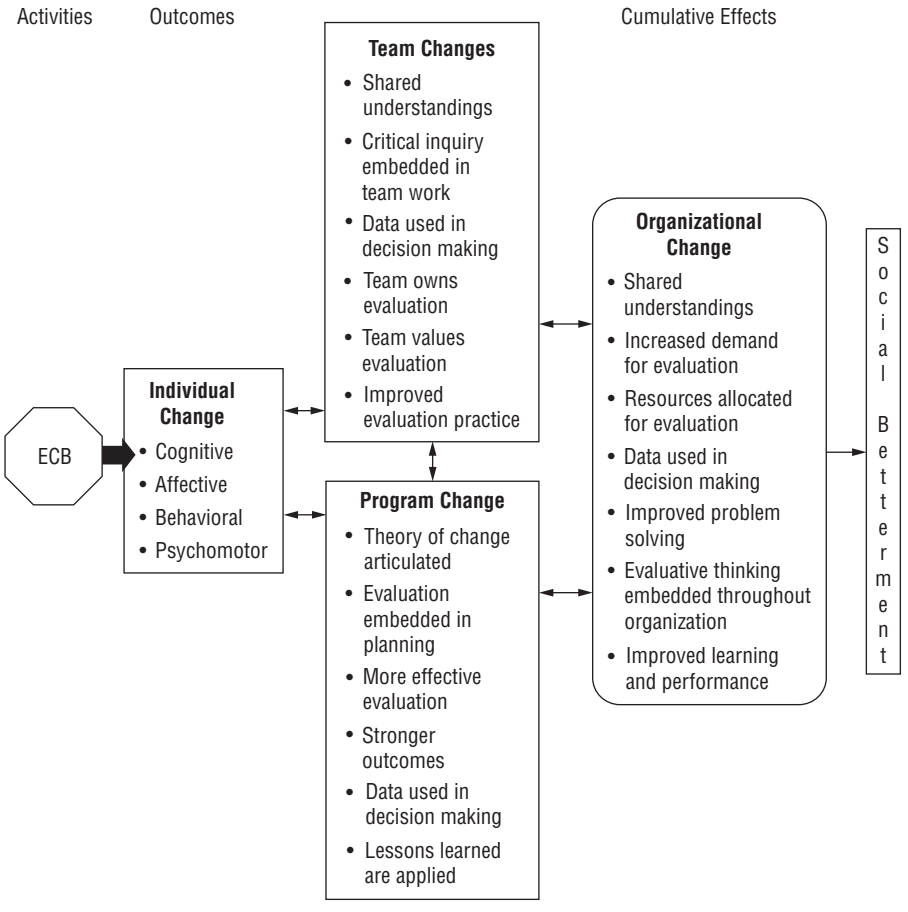


FIGURE 12.2 ***Logic Model for ECB Theory of Change***

Source: Taylor-Powell, E., & Boyd, H. H. (2008). Evaluation capacity building in complex organizations, in M. T. Braverman, M. Engle, M. E. Arnold, & R. A. Rennekamp (Eds.), Program evaluation in a complex organizational system: Lessons from cooperative extension. *New Directions for Evaluation,* No. 120, 67. This material is reproduced with permission of John Wiley & Sons, Inc.

or theories behind this logic model would have to be further articulated. For example, a program theory might be developed around the concept of social norms and how and why they would decrease teens' tobacco use. (See box on far right in model.) Such a program theory would help stakeholders to articulate the mechanisms of change: What are the social norms of teens who smoke, their friends, and family? What elements of these norms are most likely to change, for example, beliefs about health, costs, or appeal of smoking? What program components or actions will lead to changes in these norms? (The web site presents sub-logic models for parts of the program and includes narrative descriptions. See http://www.uwex.edu/ces/pdande/evaluation/pdf/YouthLMswithnarr.pdf)

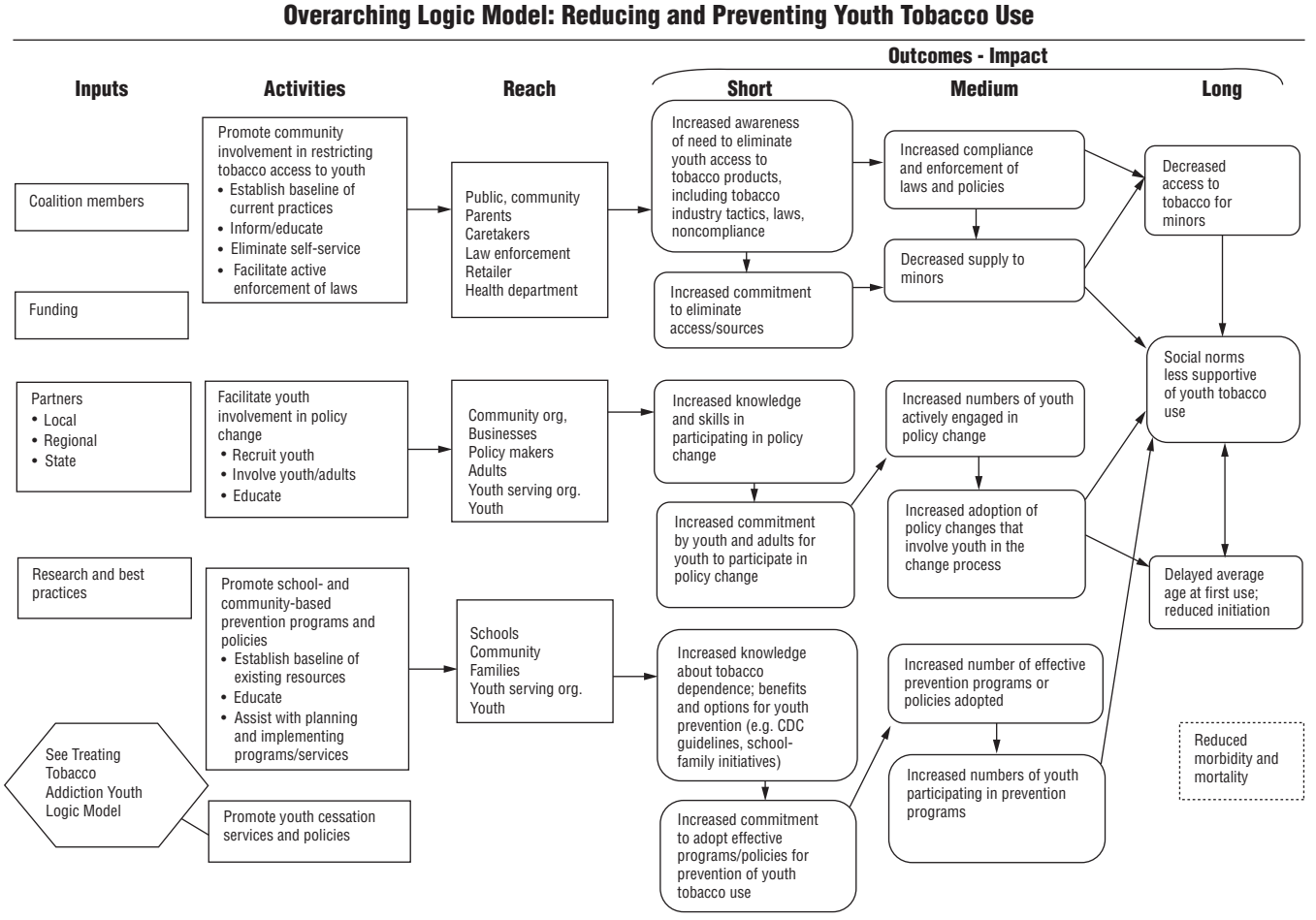


FIGURE 12.3 ***Overarching Logic Model: Reducing and Preventing Youth Tobacco Use***

Source: Taylor-Powell, E., Hrabik, L., & Leahy, J. (2005). *Documenting outcomes of tobacco control programs* (The Overarching Logic Model on Reducing and Preventing Youth Tobacco Use), Retrieved from http://www.uwex.edu/ces/pdande/evaluation/pdf/tobaccomanual. Madison, WI: University of Wisconsin Cooperative Extension.

Building Program Theory. These logic models provide a way for the evaluator to begin to describe the program, its components, and sequences. But program theory often has to be developed from the logic model if the evaluator is to thoroughly understand the causal mechanisms of the program. This emphasis is reminiscent of one of Wholey's criteria in evaluability assessment: There should be clear linkages between the program activities and outcomes that indicate why the program should have its intended effects, and these linkages should appear plausible. That is, when evaluators observe the program, review the research literature, consider the characteristics of those receiving the program, and look at the qualifications and resources available to those delivering it and the actions they perform, they should conclude that it is possible for the program goals or outcomes to be achieved if the program is to be considered evaluable. Some logic models fail to clearly articulate those linkages or what we would call the program theory. Weiss (1997) warns that the real challenges in using program theory in evaluation involve specifying the causal linkages and improving the quality of theory.

To illustrate how theory might be used to articulate causal mechanisms, we will build on a relatively simple theory first briefly proposed by Peter Rossi long ago (1971). His model suggested that the evaluator or program developer should articulate three steps: a causal hypothesis, an intervention hypothesis, and an action hypothesis.

1. The causal hypothesis links the problem to be solved or reduced by the program (A) to a purported cause (B). For example: An insufficient number of fourth-graders in our school do not succeed on the state standards test (A) because they do not know how to diagnose word problems to identify the key elements and the mathematical procedure required (B). Although a model is introduced to them, they do not spend sufficient time practicing it to become competent and do not become proficient in using it on many different types of problems (B).

2. The intervention hypothesis links program actions (C) to the purported cause (B). Therefore: Fourth-graders in our school will be introduced to the model in September and then will use it twice a week on word problems that address new math content and math taught earlier in the year (C). As a result, they will become competent at applying the model to many different types of problems by April when the state standards test is administered (B).

3. Finally, the action hypothesis links the program activities (C) with the reduction of the original problem (A). Therefore, the action hypothesis would be: Using the word problem strategy cumulatively and weekly throughout the fourth-grade year (C) will result in more fourth-grade students in our school scoring at the "competent" or above level on the state standards test on fourth-grade math.

A second example:

1. Students who graduate from our high school do not go on to postsecondary education to the extent that we would like (A) because they do not know others

(parents, friends) who have attended college, are unfamiliar with the environment, and think they will be isolated and will not succeed (B).

2. Our program that allows high school seniors to take college courses tuition free at an adjacent community college while they are still in high school (C) will help students learn about the environment, and recognize that there are other students there like themselves and that they can learn the content (B).

3. Our College during Senior Year program (C) will increase the number of our graduates who go on to postsecondary schools (A).

Note that, although simple, these models of program theory identify the reasons why the program activities may lead to a reduction of the original problem. As such, they are useful devices for discussing and articulating program theory. Through discussion, these models may become more complex; some programs will have several intervention models to illustrate their theories. Others may remain focused on one or two primary causal factors.

The evaluator can work with stakeholders to develop an impact model that specifies problems, causes, and program activities and the links between them. The initial impact model will be primarily normative, based on the perceptions and experiences of the stakeholders and program developers. But now we must attend to Weiss's second warning: develop better program theories. She writes:

> Evaluators are currently making do with the assumptions that they are able to elicit from program planners and practitioners or with the logical reasoning that they bring to the table. Many of these theories are elementary, simplistic, partial, or even outright wrong. Evaluators need to look to the social sciences, including social psychology, economics, and organization studies, for clues to more valid formulations, and they have to become better versed in theory development themselves. Better theories are important to evaluators as the backbone for their studies. Better theories are even more essential for program designers, so that social interventions have a greater likelihood of achieving the kind of society we hope for in the twenty-first century. (1997, p. 78)

As Weiss suggests, working with program staff to develop a program theory or logic model that articulates the causal mechanisms for the program is not all that the evaluator will do. The evaluator will also conduct literature searches to determine whether the causal assumptions of the developers, managers, or staff affiliated with the program are, in fact, supported by research, or whether some assumptions are questionable. In some cases, these stakeholders may have previously identified research that supports their normative model. In other cases, the evaluator might conduct a literature review to find research that supports, or challenges, the assumptions of the program model. If existing research causes questions to be raised about the existing model, the evaluator should work with program personnel and other stakeholders to revise the model. These actions to improve the program theory and achieve consensus on it, while part of the evaluation planning process, can provide immediate benefits to the program before the evaluation has even begun.

The final program theory or logic model then provides the framework for the subsequent evaluation. Thus, a causal hypothesis can lay the foundation for needs assessment questions such as: What proportion of our fourth-graders fail to achieve competence on the state standards test? How do our scores compare with other similar schools? How does our math curriculum compare with theirs? What types of items or math standards do our students tend to fail? What are the characteristics of these students who fail (new to our school, scores on other standards, performance in school)? The intervention hypothesis or hypotheses can help identify important components of the program to monitor or describe in formative studies. How much time does each class spend on the word problem-solving strategy each week? What do the teachers and students do during that time—demonstrate, practice, give feedback? What proportion of students are able to master the problem-solving model when it is first introduced? What proportion are able to continue to master it as they use it with new mathematical procedures? The action hypothesis makes the final link between the problem and the program. It exemplifies the question many black-box studies address: namely, did the program achieve its goal? However, without the important causal and intervention hypotheses, the evaluator may fail to understand why the action hypothesis has failed or is confirmed.

Methods for Describing the Program and Developing Program Theory

Developing a program description, model, or theory can be accomplished in a variety of ways. The four basic steps to collecting the necessary information are (1) reading documents with information about the program, (2) talking with various individuals who are familiar with the program, (3) observing the program in action, and (4) identifying research on elements critical to the program and its theory. Each is discussed briefly here.

Descriptive Documents. Most programs are described in proposals to funding agencies, planning documents, reports, minutes of relevant meetings, correspondence, publications, and so on. Taking time to locate and peruse such documents is an important step in understanding any entity well enough to describe it correctly.

Interviews. Helpful as they are, written documents cannot provide a complete or adequate basis for describing the object of the evaluation. It is relatively common for a program to have changed, intentionally or unintentionally, from the plans on paper to the actual implementation in the field. The evaluator should talk at length with those involved in planning or delivering the program, with people who are recipients of the program, and with those who may have observed it in operation. Stakeholders with different perspectives should be interviewed. In evaluating a treatment program for domestic violence perpetrators, for example, the evaluator

would be well advised to learn how the program is (and is supposed to be) operating, not only from the therapists and administrators responsible for delivering the program, but also from the state department responsible for providing funding for the program, the participants in the program and their families, judges who make referrals to the program, and so on. It is important to interview representatives of all the relevant audiences to develop a model on which consensus can be reached and to understand the different perspectives of the audiences.

Observations. Much can be learned by observing programs in action. In addition to personally observing the program, the evaluator may wish to ask experts on the program curricula or content to make observations. Often observations will reveal variations between how the program is running and how it is intended to run that an evaluator may not discover through interviews or reading. In fact, differences among written documents, interviews, and observations can provide the foundation for much learning at this stage. The evaluator should be alert to differences and attempt to learn how these differences have emerged. For example, when observations indicate that a program is being delivered in a way that differs from written documents and information gained from interviews, a first useful step is talking with the deliverer to learn the rationale for changes. If appropriate, these differences can then be confirmed with managers or policymakers who have been previously interviewed.

Research. The evaluator should take the time to become familiar with research on several issues: the students or clients being served, other similar programs or interventions, and the constructs that emerge as central to program theory. As noted earlier, the evaluator should be looking for whether research confirms the assumptions made in program theory or raises questions about their feasibility. (See the Fitzpatrick and Bledsoe [2007] interview for Bledsoe's work with clients and her use of research to develop a program theory.)

Finalizing the Description and Theory. Having developed the program theory, the evaluator must ensure that the stakeholders agree that it and other elements of the program description accurately characterize the program and its assumptions. Confirmation of this agreement may be achieved through ongoing meetings with a working or advisory group that has been involved in developing the program description (as in evaluability assessment), through distribution of a formal description to different audiences, or through separate meetings with various stakeholders. Recall, however, that the purpose of this stage is also to set boundaries for the evaluation by clarifying exactly what is to be evaluated. To achieve a full understanding of the context of the evaluation, the description may have involved a larger portion of the program than the evaluation will address. Now is the time to learn more from stakeholders about which elements are of interest to them and what they would like to know about those elements. Evaluators can play an active role, too, in indicating areas that they think should be studied, perhaps because of differences between research and the practitioner's theory, because of differences

between the program theory and observations of the program in action, or because of ambiguities or unknowns in the program theory.

We believe that evaluators can use these three sources and knowledge of logic models and program theory to work with stakeholders, or alone, to develop program descriptions using logic models and program theory. However, some have called for clearer methods. Leeuw writes, "How do we know what the 'nature' of the underlying theory is if our reconstruction methodology [for developing program theory] is vague?" (2003, p. 6). He identifies and discusses three different methodologies to use to "reconstruct" or create program theory and provides examples of the use of each. One method, the policy-scientific method, relies heavily on literature and research. The second, a strategic assessment approach, makes use of dialogue and consensus building with groups. The third, an elicitation methodology, builds on individuals' mental models of the program. For readers who are interested in more discussion of how to build and improve program theory, his article is useful.

Dealing with Different Perceptions

The previous discussion assumes that, in general, consensus on both the program itself and the boundaries of the evaluation exists. Such is not always the case.

As a case in point, Vroom, Colombo, and Nahan (1994) describe how differences in perceptions of goals and program priorities among managers, staff, and sponsors led to a very problematic evaluation. In an innovative program to use cable technology to help the unemployed find jobs, these stakeholders differed in the priority they attached to the technology and the direct service components of the program. The sponsor was concerned with measuring the direct impact on the unemployed, but the agency staff members were concerned with implementing the new, sophisticated cable technology. Although other organizational problems also contributed to the failure of the evaluation, the authors believe that more extensive ongoing discussions and meetings with the stakeholders could have clarified the differences in perspective and helped the evaluation.

Similarly, Donaldson describes tensions between the desire to replicate a program model that had succeeded in placing unemployed workers in Michigan in new and different settings in California and adapting the model to the exigencies of the labor market and the unemployed in California (Fitzpatrick & Donaldson, 2002). Fetterman describes the different goals, perspectives, and priorities of faculty in the Stanford Teacher Education Program that influenced his evaluation (Fitzpatrick & Fetterman, 2000).

When disagreements over the nature of the program or policy to be evaluated exist, evaluators can take one of two routes. If the differences are relatively minor and reflect the values or position of the stakeholder, evaluators can choose to learn from these different perceptions, but not push toward consensus. The different perceptions can provide evaluators with an opportunity to learn more about each stakeholder and the program if they take the opportunity to carefully learn about each group's perception or interpretation. Then, by permitting each group to attach whatever meaning they wish to the program and following

through by focusing on results that are relevant to that meaning, evaluators can address the information needs of multiple audiences. Moreover, they can educate audiences by helping them look beyond their particular perspectives.

However, if the differences of perception are major and occur among the primary audiences for the study, evaluators should attempt to achieve some sort of consensus description before moving on with the evaluation. They may want to establish a working group made up of members of the differing audiences to reach an agreement on a program description, logic model, or theory and the boundaries of the evaluation. If consensus cannot be achieved, the evaluator may conclude that further evaluation should be delayed. (See Chapter 7 for a discussion of evaluability assessment.)

Sometimes it is important for evaluators to obtain formal agreement from the client that a description is accurate. Such agreements can help avoid later conflicts between stakeholders and the evaluator about whether or not the evaluator has really understood the program to be evaluated.

Re-describing the Program as It Changes

It is important to portray the actual character of the program, not only as it begins but also as it unfolds. A critical point for evaluators to remember is that the program being evaluated frequently changes during evaluation. As House (1993) has written, a program is not "a fixed machine." The nature of a program varies and this variation is caused by many factors (McClintock, 1987). The changes may be due in part to the responsiveness of program managers to feedback that suggests useful refinements and modifications. Often a program—whether a curriculum, a training program, provision of a service, or institution of a new policy—is not implemented by users in quite the way its designers envisioned. Some adaptations may be justifiable on theoretical grounds, some may result from naïveté or misunderstanding, and some may stem from purposeful resistance on the part of users determined to expunge something objectionable from the original conception. Regardless, at the end of the evaluation, the evaluator must describe what was actually evaluated, and that may be quite different from what was originally planned.

Guba and Lincoln (1981) provide an excellent discussion of reasons why changes in the evaluation object (which they call the "evaluand") might occur.

> The evaluator who assumes that an implemented evaluand will be substantially similar to the intended entity is either naïve or incompetent. Thus, field observations of the evaluand in use, of the setting as it actually exists, and of the conditions that actually obtain are absolutely essential.
>
> Variations in the entity, setting, and conditions can occur for a variety of reasons. In some cases, the reluctance or resistance of the actors in the situation produces unwanted changes. Adaptations to fit the evaluand to the local situation may have to be made. The simple passage of time allows the action of various historical factors to make their contribution to change. Most of all, the continuing activity of the evaluator himself, if it is taken seriously by the actors and if it produces meaningful information, will contribute to a continuously changing set of circumstances. (p. 344)

A Sample Description of an Evaluation Object

To help illustrate the key points in this section, we include a discussion of a program evaluated by one of the authors (Fitzpatrick, 1988). The program to be described is a treatment program for people convicted of driving under the influence (DUI) of alcohol. The description is organized around the first two bulleted items listed on page 291, under "Factors to Consider in Characterizing the Object of the Evaluation." Program descriptions or models can be organized in many different ways. This presentation is designed to illustrate what might be learned in regard to each factor.

The First Set of Questions. What problem was the program designed to correct? What need does the program exist to meet? Why was it initiated? What are its goals? Whom is it intended to serve?

> This particular treatment program is designed for offenders who are considered problem drinkers or incipient problem drinkers due to a number of different criteria, including number of DUI arrests, blood-alcohol level at the time of arrest, scores on a measure of alcoholism, and whether an accident was involved in the arrest. The program exists to reduce deaths and accidents due to drunk driving. As with many programs of this type, it was initiated due to public attention to this issue and recognition that some sort of cost-effective treatment might be needed for certain offenders to contain the problem. Its goals are to help offenders to recognize that they have a drinking problem and to seek further treatment. The designers of the program recognize that resources for the program are insufficient to stop problem drinkers from drinking. Therefore, the following program theory contains a more immediate, implicit goal. The theory or model of the program is (a) problem drinkers drive under the influence of alcohol because they have problems with alcohol and are not aware of the extent of their problem; (b) if these offenders are exposed to information regarding alcohol use and participate in group discussions regarding their own use, they will recognize their own alcohol problem; (c) the treatment program will refer them to places where they can receive extended therapy; (d) by receiving extended therapy, the offenders will reduce their consumption of alcohol and, hence, their frequency of driving under the influence of alcohol. A secondary competing model to achieve the goal is to get participants to use alternative means of transportation and avoid driving when they have been using alcohol. (See following.)

The Second Set of Questions. What does the program consist of? What are its major components and activities? What is its basic structure and administrative/managerial design? How does it function? What research exists to link the activities of the program and the characteristics of the clients with the desired outcomes?

> The curriculum and methods used in the treatment program were developed, outlined, and disseminated to treatment sites by the state agency that funds and administers the program. The program is 20 to 30 hours in length delivered over eight to twelve sessions. The content consists of a combination of lectures, films, group discussions, and exercises. The manual for the treatment is quite specific

about the content for each session; however, interviews with the treatment deliverers, who are independent practitioners in the alcohol abuse area, indicate that they often adapt the content based on their perceptions of the group's needs. Observations of a few programs suggested experiential activities may be more limited than expected. The theory on which the treatment is based requires a heavy emphasis on experiential methods rather than didactic approaches to achieve the goal of recognizing their own problems with alcohol. Offenders who meet the criteria to be classified as problem drinkers are sentenced by the judges to complete the program at the site closest to their home. (Offenders have lost their driver's licenses as part of their punishment.) If they fail to complete the program, they are then sentenced to time in jail.

Describing Program Stakeholders. Another critical component of the program description is characterization of the stakeholders.

> Prominent stakeholders in this program include the judges (who initially sponsored the study), Mothers Against Drunk Driving (MADD, who lobbied the judges to sponsor the study), the program deliverers, and the state division that oversees the program. Other stakeholders include the clients, their families, victims of traffic accidents in which arrests are involved and their families, insurance companies, alcohol treatment centers, and the public at large. The judges are interested in improving their sentencing by learning what kinds of offenders are least likely to complete treatment successfully. MADD is interested in the degree to which the program achieves its goals. The program deliverers are persons with expertise in treating alcohol offenders both through these programs and others. They are most interested in ending alcohol abuse. The state funding source agrees that alcohol abuse is part of the problem and should be remedied but also advocates alternative methods of transportation. Research suggests that treatments that decrease the frequency of driving (removal of license and increase in insurance costs) are more effective at decreasing deaths due to driving under the influence than treatments that address the alcohol abuse itself. However, as most professionals in this field are from the alcohol treatment area rather than the transportation area, the alcoholism treatment approach tends to be the dominant focus.
>
> Much research has been conducted in this area nationally, and the state division responsible for overseeing the program collects some routine data from sites. However, little systematic research has been performed on this state program. Current data collection focuses on attendance and change on pre-post knowledge and attitude measures administered at the beginning and end of the programs. No monitoring studies or follow-up outcome studies have been conducted.

The preceding description is designed to illustrate some of the critical factors that an evaluator might note in characterizing the object of the evaluation during the planning phase. To describe the program, the evaluator made use of printed material (state manuals and proposals, local site materials); interviews (judges, MADD, deliverers, state administrators); observations of the program; and a review of literature on treatment programs and solutions for people convicted of drunk driving. With the information obtained from answering the first two questions,

a model can be developed for the program. The model can be depicted with a flowchart or in a narrative manner, as illustrated in this example. Empirical research that both supports and weakens the model should be described. Site visits to the program can help bring this model to life and assist the evaluator in determining whether program activities correspond to the model. This description, then, can provide the foundation for further communication with audiences in planning the focus of the evaluation.

Analyzing the Resources and Capabilities That Can Be Committed to the Evaluation

Very often, program managers, deliverers, and, sometimes, clients themselves view resources committed to evaluation as resources taken away from the program itself. They believe that if only those dollars were available, they could educate more students, treat more patients, serve more clients, create more parks, and so on. However, others in the public and nonprofit sectors have come to recognize that evaluation can be very useful to them. Evaluation can help program managers and deliverers adapt programs to better meet the needs of their clients. Hodgkinson, Hurst, and Levine (1975) first introduced the doctrine of cost-free evaluation to argue that evaluation is not an "added-on-extra" but a means for identifying "cost-saving and/or effectiveness-increasing consequences for the project" (p. 189), just as private-sector companies make use of research on their products and modes of delivery. Evaluators should recognize that their purpose is to improve productivity and the quality of the product, either through formative recommendations for program improvement that will lead to better products or result in lower costs or through summative recommendations that will result in maintaining or expanding successful, cost-effective programs or eliminating unsuccessful ones.

Analyzing Financial Resources Needed for the Evaluation

Even when the client is converted to the doctrine of cost-free evaluation, determining what resources can be devoted to evaluation is difficult. As Cronbach and others (1980) noted, "deciding on a suitable level of expenditure is . . . one of the subtlest aspects of evaluation planning" (p. 265). In Chapter 14, we will make specific suggestions about managing an evaluation and developing a budget. Here, during this early stage in planning, we will highlight some issues the evaluator should consider so that the scope of the evaluation does not become too large for existing resources and capabilities.

Ideally, the decision about the resources that will be made available for the evaluation should be made in consultation with the evaluator, whose more intimate knowledge of evaluation costs would be of great help. Unfortunately, there may not be sufficient rapport between evaluator and client to foster such collaborative

planning, though such collaboration is much more likely to occur when the evaluation is conducted by internal evaluation personnel. In many situations in which external evaluators are hired, the client may initially proceed independently to set budgetary limits for the study. Sometimes the evaluator is informed about how much money is available for the evaluation. Frequently, however, this is not made clear. In such cases, we recommend that the evaluator propose two or three different levels of evaluation that differ in cost and comprehensiveness—perhaps a "Chevrolet" and a "Cadillac" evaluation, for example—from which the client can select. Internal evaluators can have more dialogue with decision makers at the budgeting stage, but may also develop several budgets to reflect different alternatives. Clients new to evaluation are often unaware of the possibilities of evaluation design, of what information evaluations might be able to produce, or of the cost of evaluation services. Faced with decisions about tradeoffs and budget limitations, the client also needs to know about alternatives and their consequences to make a good decision. Budgeting could be the last step in planning an evaluation. In contrast, if budget limits are known at the beginning, they will affect (and usually enhance) planning decisions that follow.

Evaluation plans and budgets should remain flexible, if at all possible. Circumstances will change during the study, and new information needs and opportunities will unfold. If every dollar and every hour of time are committed to an inflexible plan, the results will fail to capitalize on the new insights gained by evaluator and client. Even the most rigid plan and budget should include provisions for how resources might be shifted, given approval of the client or decision maker, to accomplish evaluation tasks that take on new priority through changing circumstances.

Analyzing Availability and Capability of Evaluation Personnel

Budget is only one consideration affecting the design of an evaluation study. Personnel is another. Both internal and external evaluators can be assisted by employees whose primary responsibility is to perform other functions. Program deliverers may be able to collect data. Administrative assistants can prepare documents, search records, and make arrangements for meetings, interviews, or the like, at no cost to the evaluation budget.[1] Graduate students from local universities seeking internship experience or working on dissertations or course-related studies can undertake special assignments at minimal cost to the evaluation budget. Volunteers from neighborhood associations and other community groups, parent–teacher associations, church groups, or advocacy groups associated with the project can often perform nontechnical evaluation tasks. Clients themselves can also help. Calls for volunteers from among these various groups often pay off, and involving volunteers not only helps contain costs but also sparks interest in the evaluation among stakeholders.

[1]This is not to say there is no cost for such personnel services, only that it may be possible to obtain some assistance with evaluation tasks from on-site personnel within existing operating budgets.

Whenever people who are not evaluation specialists conduct or assist with evaluation tasks, the evaluator faces unique responsibilities that cannot be neglected: orientation, training, and quality control. Evaluation personnel who lack specialized training or relevant experience require orientation to the nature of the study, its purposes, and the role they will be asked to play. They must understand their responsibilities, not only in completing evaluation tasks in an effective and timely manner, but also in representing the evaluation team and its sponsoring organization. Naïve and unprepared evaluation staff (volunteers or otherwise) can play havoc with an evaluation if they interact abrasively with others, misrepresent the nature or purpose of the study, betray anonymity or confidentiality, or even dress inappropriately for the setting. Volunteers or assistants must also be trained in the skills required to do the tasks assigned. They must follow protocols or misleading, inaccurate information or other types of errors may result. Supervision and spot checking can be very helpful, especially in the early stages, to ensure that nonevaluation personnel understand their tasks and responsibilities.

Using nonevaluation personnel to expand an evaluation effort at low cost also introduces the risk of bias. Personal considerations must not influence the way in which these volunteers conduct their evaluation tasks. It is easy to allow presuppositions to color one's perceptions. Evaluators are trained to recognize their biases and adjust and, even with that training, it is a struggle. People new to evaluation are unlikely to be aware of their biases or the effects they may have on the evaluation. Although few seem likely to be so unprincipled, it is also possible to alter or distort the data to make it fit one's prior conclusions. Thus, to protect the study's validity and credibility, it is essential that an evaluator exercise caution and judgment in determining which tasks the nonevaluation volunteers will perform. If they are assigned tasks that would expose them to confidential information, they should be carefully trained in the meaning of confidentiality and the importance of maintaining the dignity and privacy of clients or others from whom data are collected. Given conscientious supervision, monitoring, and auditing, local staff or volunteers can make a valuable, cost-effective contribution to an evaluation.

Some evaluations have been greatly enhanced by using people not originally trained in evaluation to assist in a project. Mueller (1998) won an American Evaluation Association award for the work her team did in evaluating Minnesota's Early Childhood Family Education (ECFE) program. Part of the impetus for the award was the effective training and use of program staff at all stages of the evaluation. Staff were involved in identifying families for different phases of the study, videotaping and interviewing clients, analyzing data, and developing reports. She was able to conduct a relatively comprehensive evaluation with a small budget and to achieve her goal of building internal evaluation capacity within the organization by training and using program staff to perform the evaluation. Given appropriate training, volunteers who are part of the client population can sometimes collect more sensitive and valuable qualitative information through interviews than well-trained but culturally different evaluators. Empowerment evaluation is founded on the principle of training others to evaluate their own programs. (See Christie and Conner [2005] for an interview in which Conner describes his use of community residents in a state-wide evaluation.

Other interviews that illustrate the use of program deliverers or citizens in data collection include Wallis, Dukay, & Fitzpatrick, 2008; and Fitzpatrick & King, 2009).

Analyzing Technological and Other Resources and Constraints for Evaluations

The availability of existing data, including files, records, previous evaluations, documents, or results of other data-collection efforts to which the evaluation may be attached, is an important consideration. The more information that must be generated by the evaluator, the more costly the evaluation.

The availability of needed support materials and services is also important. Existing testing programs, computer services, routine questionnaires, or other information services are all possible resources that could be drawn on at little or no cost to the evaluation if they already exist for other purposes.

Advances in technology have provided opportunities both for collecting more useful information and for reducing costs. Face-to-face communication, particularly with new stakeholders during the planning stage, can never be replaced. However, e-mail among groups and individuals can be used to increase communication and reduce time and travel costs by replacing or supplementing meetings involving people from different locations, sometimes scattered across a state or country. Conference calls, with video accompaniment, can sustain the dynamics of face-to-face meetings while reducing costs. Technology can be used to share drafts of data-collection measures, results, or reports and to seek input from various stakeholders who might not have been included in the past. Surveys can be conducted with target audiences who are online and results can be analyzed as data are accumulated. Videos or photographs of program activities, stakeholders, clients, or other pertinent evaluation- or program-related information can be posted on the Internet for others to view and comment. Finally, it has become commonplace for many final evaluation reports to be posted on web sites for foundations or organizations so the results can reach wider audiences. (See Fitzpatrick and Fetterman [2000] for a discussion of Fetterman's use of technology in the evaluation of the Stanford University teacher-training program.)

Time must be considered a resource. The evaluator does not want to miss opportunities for making the evaluation useful because of tardy reports, or data collection and analysis. Knowing when to be ready with results is part of good planning. Ideally, an evaluator would have sufficient time to meet all information needs at a pace that is both comfortable and productive. Limited time can diminish an evaluation's effectiveness as much as limited dollars.

Analyzing the Political Context for the Evaluation

Evaluation is inherently a political process. Any activity that involves applying the diverse values of multiple constituents in judging the value of some object has political overtones. Whenever resources are redistributed or priorities are redefined, political processes are at work. Consider the political nature of decisions regarding whose

values are attended to, how they are weighted, what variables are studied, how information is reported and to whom, how clients and other audiences intend to use evaluative information, what kind of support is given to the evaluation and by whom, what potentially embarrassing information is hidden, what possible actions might be taken to subvert the evaluation, and how the evaluator might be co-opted by individuals or groups. Political processes begin to work with the first inspiration to conduct an evaluation and are pivotal in determining the purpose(s) to be served and the interests and needs to be addressed. Political considerations permeate every facet of evaluation, from planning to the reporting and use of evaluation results.

We discussed ways to work with political factors in evaluation in Chapter 3, but we cannot leave this chapter without saying a few words about the importance of analyzing the political context in which the evaluation will be conducted while there is still time to recognize and retreat from a political debacle that could render an evaluation useless.

On receiving any new request to undertake an evaluation, the evaluator might consider the following questions:

1. Who would stand to lose/gain most from the evaluation under different scenarios? Have they agreed to cooperate? Do they understand the organizational consequences of an evaluation?

2. Which individuals and groups have power in this setting? Have they agreed to sanction the evaluation? To cooperate?

3. How is the evaluator expected to relate to different individuals or groups? As an impartial outsider? An advocate? An organizational consultant? A future consultant or subcontractor? A confidante? A facilitator? What implications does this have for the evaluation and its ability to provide useful results in an ethical manner?

4. From which stakeholders will cooperation be essential? Have they agreed to provide full cooperation? To allow access to necessary data?

5. Which stakeholders have a vested interest in the outcomes of the evaluation? What steps will be taken to give their perspective a fair hearing without allowing them to preclude alternative views?

6. Who will need to be informed during the evaluation about plans, procedures, progress, and findings?

7. What safeguards should be incorporated into a formal agreement for the evaluation (reporting procedures, editing rights, protection of human subjects, access to data, metaevaluation, procedures for resolving conflicts)?

Answers to these questions will help the evaluator determine whether it will be feasible and productive to undertake the evaluation study—a decision we will address shortly. First, it may be helpful to consider briefly how the activities and issues discussed so far in this chapter would be influenced by the evaluation approach being used.

Variations Caused by the Evaluation Approach Used

The participant-oriented model has had an important influence on evaluators. Few evaluators today would conduct an evaluation without considering the perceptions and needs of other stakeholders and the context in which the evaluation is to be conducted. However, as in Chapter 11, the models differ in emphasis.

Evaluators using a pure objectives-oriented approach, a rare case today among trained evaluators, might involve different audiences in defining program objectives but, in their single-minded focus on objectives, might fail to obtain an adequate description of the program and an understanding of the political context in which it operates. An objectives-oriented approach tends to be relatively linear and can fail to acknowledge the multiplicity of views about the program, the clients it serves, and the context in which it operates. However, the development of logic models or the specification of program theory, as advocated by those using a contemporary program-oriented approach and as discussed in this chapter, can avoid that problem. The development of program theory or logic models, especially when conducted as a dialogue with different stakeholders, can illuminate program operations and hidden assumptions.

Similarly, a decision-oriented approach is often criticized for its focus on managers as the primary decision makers and for providing information only for the identified decisions to be made. While sophisticated users of this model would certainly identify and learn about the concerns of other stakeholders, these groups would be viewed as secondary. If such stakeholders were outside the organization (e.g., clients, interest groups, elected officials), they would almost certainly not be seen as a primary audience. The evaluator following this approach would tend to see them as lacking the power to make decisions that could affect the program dramatically. (Obviously, such evaluators would have failed to consider effective political actions at the grassroots level or the power of the school board or legislature!) Similarly, a decision-oriented evaluator might focus on defining the decisions to be made and the context for those decisions rather than on the context for the program itself. Today's performance monitoring and use of standards falls within the decision-oriented model because these devices are used for internal management decisions. However, the decisions to be made often are not specified and, like early objectives-oriented models, the evaluation may provide no information for how the standard or desired level of performance might be achieved.

The consumer-oriented approach will, of necessity, define the program from the perspective of consumers. In this case, other audiences and other views of the program or products may be neglected. Thus, a consumer-oriented evaluation of the national forests might choose to focus on the satisfaction of campers in these forests. How pleased are they with the camping facilities? The beauty of the site? The access to the campground? Such a focus would neglect other audiences, such as ranchers, nonusers who want the land protected, and future generations of users and nonusers. The expertise-oriented evaluator is likely to be the most narrow in identifying and considering stakeholders and their descriptions and views of the program. They are hired more for their expertise in the content of the program than for

their expertise in evaluation. Such knowledge, and the criteria for the evaluation, typically arise from professional education, training, experience in the field, and, often, standards developed by the same profession in which the "expert" is educated. Thus, the stakeholders for the program and the means for describing the program are rather narrowly circumscribed by the profession the program represents (e.g., math educators for math curricula, medical personnel for hospitals, criminal justice for prisons). The expertise-oriented evaluator may collect data on the program from many different stakeholders but rarely would consider these stakeholders' and audiences' information needs for the evaluation. These evaluators would view their role as reflecting the standards of experts in the field, not those of others.

The participant-oriented model is certainly the most ardent in advocating the inclusion of many different stakeholders and their perspectives in the planning of the evaluation. Evaluators using this model would seek the multiple perspectives of different stakeholders, arguing that no one view of the program reflects absolute truth and, therefore, they must seek many different perspectives to describe the program and consider the purposes of the evaluation at this stage. Of course, some participatory approaches involve only a few stakeholders, but in great depth; such evaluators would need to identify and prepare these stakeholders for their roles. Others involve many stakeholders but in less depth. The question becomes one of synthesizing the program description and the purposes of the evaluation. Who makes this synthesis? The evaluators may become important decision makers with this approach. But, in even the most evaluator-led participatory approaches, much stakeholder input is sought in the planning stage. Decision-oriented evaluators might accuse the participant-oriented evaluators of being naïve about the political sphere, because they are betting that the managers or primary users they target are the ones who are most interested and most able to make decisions based on the results of the evaluation. Participant-oriented evaluators might retort that few decisions emerge directly from an evaluation. By involving and informing many audiences, their evaluations, they would argue, are more likely to make a difference in the long run.

Determining Whether to Proceed with the Evaluation

In Chapter 11, we talked about identifying reasons for the evaluation. Such reasons provide the best indicators of whether an evaluation will be meaningful. In this chapter we have discussed the importance of understanding who will use the evaluation information and how, and we have suggested ways to identify relevant audiences. We have stressed the importance of describing and setting boundaries for what is evaluated and analyzing fiscal, human, technological, and other resources to determine feasibility. We have cautioned evaluators to consider whether any political influences might undermine the evaluation effort as well.

At this point, the evaluator must make a final determination—having accumulated sufficient information on the context, program, stakeholders, and resources available—of whether to continue the evaluation. In Chapter 11, we reviewed conditions under which an evaluation might be inappropriate. Now, the evaluator has learned even more about the boundaries and feasibility of the evaluation and the

needs and views of stakeholders. With this information, the evaluator should again consider whether to proceed. Unfortunately, we can offer no simple algorithm for balancing all these factors to make a final decision about whether to proceed with the evaluation. Thoroughness in considering the factors outlined in this and the preceding chapter—insight, thoughtfulness, and common sense—are the ingredients essential to a sensible decision about when to agree to do an evaluation. Daniel Stufflebeam has developed a checklist that is useful in considering whether you have addressed all the important factors involved in planning an evaluation. It can be seen at http://www.wmich.edu/evalctr/checklists/plans_operations.pdf

Although we have painted this as a yes or no decision, other options, of course, exist. Having learned more details about the context, the program, and the resources available, evaluators can work with the client or sponsor to limit the evaluation to an area of the program that may be most fruitful and useful to evaluate, or to a more narrow scope.

Major Concepts and Theories

1. A first step in analyzing the evaluation context is learning the needs and perceptions of the evaluation from different stakeholders. Identify, interview, and, as appropriate, involve different potential users in the planning stage.

2. The second means for setting boundaries and understanding the context of the evaluation is developing a program description, which can include a logic model or program theory. The larger description and model or theory can be developed through interviews with stakeholders or as a group, but should ultimately be shared with the users for confirmation or discussion.

3. Logic models describe the process of the program—its inputs, activities, outputs, and outcomes. Some logic models can serve as important tools for guiding implementation studies and can give some idea about assumed causal linkages, that is, links between the initial problem, program activities, and goals.

4. Understanding the theory of the program is essential at this stage. Program theory specifies the connections between the problem(s) to be resolved, program actions, and program goals. It can serve as a foundation for the evaluation questions and to acquaint the evaluator with the very essence of the program. Program theory can include stakeholders—particularly the views of program developers, managers, and deliverers as to why the program should work—but should ultimately be adjusted to research findings and theories on the issue.

5. To describe the program fully and to develop logic models or program theory, the evaluator should review existing information (e.g., organizational reports, proposals, previous evaluations); interview managers, staff, clients, and other important stakeholders; review literature in areas related to the program; and observe the program in operation.

6. Consider available resources and the potential costs associated with the evaluation. Program staff or volunteers may be used to reduce costs.

7. Consider how the political context may affect the approach of the evaluation, the nature of information collected, and the interpretation and use of results

Discussion Questions

1. Why is it important to consider all the different stakeholders and audiences for an evaluation? Which groups do you think are typically viewed as most important? Which are most likely to be neglected? How could ignoring these latter groups lead to a problem?
2. Why is it important to understand the context of the object being evaluated?
3. What role do you think research on program-related issues plays in understanding program theory?
4. What are some of the advantages and risks in using program staff to assist in an evaluation? In using volunteers?
5. Why is evaluation inherently a political process?

Application Exercises

1. Consider a program with which you are familiar. Who are the stakeholders for this program? What additional audiences exist for the results of the evaluation? Use Figure 12.1 to identify potential stakeholders. Whom might you choose to interview? What might be the perspectives of each? Do you think it would be advisable to select a representative of each stakeholder group to serve as an advisory committee for the evaluation? If so, whom would you select and why? Would you omit certain stakeholder groups or individuals? Why?
2. What critical political factors might the evaluator in Exercise 1 need to be aware of?
3. Through either a literature review or your own workplace, find a report or brochure describing a program. (You might examine the logic models posted at http://www.uwex.edu/ces/pdande/evaluation/evallogicmodelexamples.html or consider a program you know in your workplace.) What are the goals and objectives for this program? What are the critical components and activities? Is it feasible that these goals and objectives could be achieved with the specified clients using the described activities? Why or why not? What is the program theory? Does the literature review or program description provide any evidence for why the model should work? For why it might fail? Does it provide an accurate description of the model? What questions would you like to ask program staff to learn more about the model?
4. Consider the problem of teacher turnover or employee turnover in your agency. Develop an impact model for this problem with a causal hypothesis, intervention hypothesis, and action hypothesis. (Hint: It would begin: Employees leave our organization because . . . .) First, develop such a model based on your knowledge of the problem. Next, interview people in your community, school, or agency and develop a normative theory of the problem. Finally, review the research literature and determine the validity of the normative model based on the research you find. What alternative causative models might you develop?

Relevant Evaluation Standards

We consider the following evaluation standards, which are listed in their entirety in Appendix A, to be relevant to this chapter's content:

U2—Attention to Stakeholders
U3—Negotiated Purposes
U4—Explicit Values
U5—Relevant Information
F3—Contextual Viability
F4—Resource Use
A4—Explicit Program and Context Descriptions
E1—Evaluation Documentation

Case Studies

For this chapter, we recommend three interviews that demonstrate different ways for setting the boundaries and analyzing the context for the evaluation: *Evaluation in Action*, Chapters 9 (Stewart Donaldson), 10 (Hallie Preskill), and 12 (Katrina Bledsoe).

In Chapter 9, Stewart Donaldson describes his dialogue with stakeholders, including the sponsor, clients, program developers, and others, to focus the evaluation of four different programs over time. One of his primary tools, but not the only one, is developing and using the theory for the program. The journal source is Fitzpatrick, J. L., & Donaldson, S. I. (2002). Evaluation of the Work and Health Initiative: A dialogue with Stewart Donaldson. *American Journal of Evaluation*, *23* (3), 347–365.

In Chapter 10, Hallie Preskill discusses her use of Appreciative Inquiry with an organization to help them build an evaluation system. The chapter illustrates how, in a participatory, capacity-building way, Preskill leads the participants to consider their preferences in an evaluation system. The journal source is Christie, C., & Preskill, H. (2006). Appreciative Inquiry as a method for evaluation: An interview with Hallie Preskill. *American Journal of Evaluation*, *27*(4), 466–474.

In Chapter 12, Katrina Bledsoe describes how she works to learn the context of the program, develop its theory, use the research literature, and define the focus of the evaluation. The journal source is Fitzpatrick, J. L., & Bledsoe, K. (2007). Evaluation of the Fun with Books Program: A dialogue with Katrina Bledsoe. *American Journal of Evaluation, 28*(4), 522–535.

Suggested Readings

Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications.* New York: Lawrence Erlbaum and Associates.

Leeuw, F. K. (2003). Reconstructing program theory: Methods available and problems to be solved. *American Journal of Evaluation*, *24*(1), 5–20.

Mueller, M. R. (1998). The evaluation of Minnesota's Early Childhood Family Education Program: A dialogue. *American Journal of Evaluation*, *19*(1), 80–99.

Stufflebeam, D. L. (1999). Evaluation plans and operations checklist. http://www.wmich.edu/evalctr/checklists/plans_operations.pdf

United Way of America. (1996). *Measuring program outcomes: A practical approach*. Alexandria, VA: United Way of America.

Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. In D. Rog & D. Fournier (Eds.), *Progress and future directions in evaluation: Perspectives on theory, practice, and methods.* New Directions for Evaluation, No. 76. San Francisco: Jossey-Bass.

13

Identifying and Selecting the Evaluation Questions and Criteria

Orienting Questions

1. What is the function of evaluation questions? Criteria? Standards? When are criteria and standards necessary?
2. What are good sources for evaluation questions?
3. What role should the evaluator play in determining what questions will be addressed in the evaluation? What role should the client play?
4. In identifying and selecting evaluation questions, what different concerns and activities are involved in the divergent and convergent phases?
5. Should standards be absolute or relative?

Evaluations are conducted to answer questions concerning program adoption, continuation, or improvement. The evaluation questions provide the direction and foundation for the evaluation. Without them, the evaluation will lack focus, and the evaluator will have considerable difficulty explaining what will be examined, and how and why it is being examined. This chapter will focus on how these evaluation questions can be identified and specified to provide the foundation for the evaluation study and to maximize the use of the results. Evaluators' primary responsibility is to work with stakeholders and to use their own knowledge of and expertise in research and evaluation to develop questions that are meaningful, important, feasible to answer within the given resources, and likely to provide useful information to primary intended users and other stakeholders.

The process of identifying and defining the questions to be answered by the evaluation is critical. It requires careful reflection and investigation. If important questions are overlooked or trivial questions are allowed to consume evaluation resources, it could result in the following:

- Little or no payoff from the expenditure for the evaluation
- A myopic evaluation focus that misdirects future efforts
- Loss of goodwill or credibility because an audience's important questions or concerns are omitted
- Disenfranchisement of legitimate stakeholders
- Unjustified conclusions about the program

During this planning stage, evaluators may also work with stakeholders to identify the criteria, or factors, that will be used to judge the success of the program and the standards to be used to determine program success on each criterion. Identification of criteria and standards are particularly important in summative evaluations, but also play a role in formative evaluations. In both cases, the evaluation typically makes a judgment about the quality of the program—or some aspect of it: whether it should be continued or expanded or whether it should be improved discontinued. We will return to identifying criteria and setting standards later in the chapter, but first we will focus on developing and identifying the evaluation questions to guide the study.

Cronbach (1982) used the terms "divergent" and "convergent" to differentiate two phases of identifying and selecting questions for an evaluation. We will adopt these helpful labels in the discussion that follows.

In the divergent phase, as comprehensive a laundry list of potentially important questions and concerns as possible is developed. Items come from many sources, and little is excluded, because the evaluator wishes to map out the terrain as thoroughly as possible, considering all possible directions.

In the convergent phase, evaluators select from this list the most critical questions to be addressed. Criteria and standards may then be specified for questions that require them. As we shall see later in this chapter, the process of setting priorities and making decisions about the specific focus for an evaluation is a difficult and complex task.

During the evaluation, new issues, questions, and criteria may emerge. The evaluator must remain flexible, allowing modifications and additions to the evaluation plan when they seem justified. Now let us consider the divergent, and then the convergent, phase in some detail.

Identifying Useful Sources for Evaluation Questions: The Divergent Phase

Cronbach (1982) summarizes the divergent phase of planning an evaluation as follows:

> The first step is opening one's mind to questions to be entertained at least briefly as prospects for investigation. This phase constitutes an evaluative act in itself, requiring collection of data, reasoned analysis, and judgment. Very little of this

> information and analysis is quantitative. The data come from informal conversations, casual observations, and review of extant records. Naturalistic and qualitative methods are particularly suited to this work because, attending to the perceptions of participants and interested parties, they enable the evaluator to identify hopes and fears that may not yet have surfaced as policy issues.
>
> The evaluator should try to see the program through the eyes of the various sectors of the decision-making community, including the professionals who would operate the program if it is adopted and the citizens who are to be served by it (pp. 210, 212–213).

For evaluators to obtain genuinely diverse views about what the evaluation might address, they must throw a broad net and learn from many possible sources. These sources include:

1. Information needs, questions, and concerns of stakeholders
2. Questions or issues suggested by evaluation approaches (such as those in Part Two of this book)
3. Theories and findings in the research literature on the content of the program or its clients
4. Professional standards, checklists, guidelines, or criteria developed or used elsewhere
5. Views and knowledge of expert consultants
6. The evaluator's own professional judgment

Each of these sources will be discussed in more detail in the following pages.

Identifying Questions, Concerns, and Information Needs of Stakeholders

Generally, the single most important source of evaluation questions is the program's stakeholders: its clients, sponsors, participants, and affected audiences. Today, most approaches to evaluation emphasize the importance of consulting stakeholders, particularly during the planning phase. We cannot overemphasize the importance of garnering the questions, insights, perceptions, hopes, and fears of the evaluation study's stakeholders, for such information should be primary in determining the evaluation's focus.

To obtain such input, the evaluator needs to identify individuals and groups who are influenced or affected by whatever is being evaluated. The checklist of potential evaluation stakeholders and audiences presented in Chapter 12 can be used to identify potential stakeholders to involve in the identification of questions. If the approach to be taken will include involving a few stakeholders in great depth, evaluators should take the time now to identify them and begin their intensive involvement. Patton (2008a) advises that such stakeholders should be primary intended users, people or even an individual who is interested in the evaluation and motivated and able to use the results. Other approaches require the evaluator to involve many stakeholders at this stage.

It can be useful for evaluators to winnow the extensive list in Figure 12.1 to a few categories. We find that useful categories of stakeholders to consider for evaluation questions include (1) policymakers (such as legislators, legislative staff, or governing board members and their staff); (2) administrators or managers (those who direct and administer the program or entity to be evaluated or who manage the organization in which the program resides); (3) practitioners or program deliverers (those who operate the program or deliver its services); (4) primary consumers (those who participate in the program and are intended to benefit from it); and (5) secondary consumers (family members, citizens, and community groups who are affected by what happens to primary consumers or to the program itself). These five groups represent types of stakeholders who are associated with almost any program. Several distinct stakeholders or groups of stakeholders may emerge in each category. For example, administrators and managers for a school program will often include assistant principals, principals, people in the central administration who are affiliated with the program, cluster coordinators, and so on.

Once stakeholders are identified, they should be interviewed to determine what they would like to know about the object of the evaluation. What questions or concerns do they have? What are their perceptions of the program to be evaluated? What do they think it is designed to do, and how well do they think it is doing that? What do they know about the program activities, and do they have concerns about any particular elements or phases? What do they see as the rationale or reasoning for the program and how it works? How would they change the program if they had the opportunity?

Dynamics of Involving Stakeholders to Achieve Validity and Equity. Evaluation has moved in an increasingly participative direction since its beginnings. Today, stakeholders are involved in evaluations for many reasons, but the primary ones are to encourage use and to enhance the validity of the study (Brandon, 1998; Cousins & Whitmore, 1998). Participative evaluators have been persuasive in arguing that involving potential users at many stages of the evaluation will increase use of the results. Involving stakeholders in the planning phase reduces their anxiety about the evaluation and improves their understanding of its purposes and intent, as well as ensuring that the evaluation questions address their concerns.

Involving stakeholders has the further advantage of increasing the validity of the study (Brandon, 1998, 2005). Evaluators, especially external evaluators, may be new to the program; stakeholders are not. They know it. Huberman and Cox have written: "The evaluator is like a novice sailor working with yachtsmen who have sailed these institutional waters for years, and know every island, reef, and channel" (1990, p. 165). Involving stakeholders in describing the program, setting program boundaries, identifying evaluation questions, and making recommendations about data collection, analysis, and interpretation adds to the validity of the evaluation because stakeholders are program experts. While the expertise of stakeholder groups will vary, each group has a particular view of the program that is different, and often more knowledgeable, than that of the evaluator. Students or clients have experienced the program intimately as recipients. Staff have

delivered the program and often made choices about how it is delivered. Managers have helped fund and plan it, have monitored it, and have hired people to implement it. The stakeholders are program experts, but evaluators are typically the evaluation experts. They know what evaluation can do and, equally important, what it cannot do. So, communication between the two, stakeholders and evaluators, is needed to identify questions that evaluation studies can successfully address and that are meaningful and useful to the intended users.

Nick Smith (1997) has developed three broad procedural rules for using stakeholders to improve evaluation:

- Stakeholders differ in their knowledge and expertise. Use stakeholders for the areas in which they have expertise and experience.
- Consider carefully the methods used to tap that expertise.
- Make sure that participation is equitable, in particular that stakeholders with less power are able to provide information and views in a safe, comfortable, and equitable fashion.

Brandon (1998), Greene (1987), and Trochim and Linton (1986) describe some specific methods for achieving useful and valid input from stakeholders. Ask stakeholders about what they know. For example, teachers know why they made changes in the planned curricula; students or program participants do not. Participants know what they understood and how they felt about a program or curriculum; teachers or program staff may not be the best source for such information. The evaluator must consider what each group knows and learn more about that group's perspective. Brandon (1998) describes an excellent way for involving teachers—the group with the most expertise in appropriate expectations for their students—in setting standards for an evaluation.

When stakeholder groups differ in power, as is the case in almost all evaluations, using small groups, trained facilitators, and other methods to hear the voices of less powerful stakeholders can be important. In educational evaluations, many parents who have not had successful experiences with the school system (e.g., parents who struggled in school themselves, immigrants, non-English-speaking parents) are unlikely to feel comfortable expressing their concerns in a large group where teachers and educational administrators, often of a different social class, are present. Students are likely to feel similarly disenfranchised. Yet the views of these groups are important, not only for democratic and social purposes, but also to improve the validity of the evaluation itself. These groups can provide significant and different perspectives on the evaluation questions and methods of data collection, but the manner in which their input is sought must be carefully considered and planned.

Eliciting Evaluation Questions from Stakeholders. Many stakeholders who are unfamiliar with evaluation may have difficulty expressing what they would like the evaluation to do because they do not know what evaluations can do. It is, therefore, important that the evaluator collect information in ways that are meaningful to stakeholders. Rather than focus on the evaluation, the evaluator

can begin with the stakeholders' area of expertise—their knowledge of the program, their experience with it, and their concerns about it. The evaluator may translate these concerns into evaluation questions at a later point.

In many cases, as relationships with significant stakeholders evolve, the evaluator may move into an educative role to help the stakeholders learn about the different questions the evaluation could address or to acquaint them with relevant research findings or evaluation approaches that would be appropriate. However, at the initial stage, it is important for the evaluator to spend more time listening than educating. By listening to stakeholders' perceptions and concerns, the evaluator will gain an enormous amount of information about the program, its environment, typical methods of decision making, and the values and styles of the stakeholders. Asking why they are concerned about a particular aspect of the evaluation object, why they value specific outcomes, what other methods they think would be useful for achieving the outcomes, or what they would do with the answers to particular questions can help the evaluator judge the value of those questions.

There is no single technique for eliciting evaluation questions from stakeholders, but we believe a simple and direct approach works best. Before attempting to identify these questions, it is useful to establish a context that will help make them more meaningful. For example, we might begin this way: "As you know, I've been hired to do an evaluation of the X program. I would like the information that I collect to be useful to people like yourself. At this stage, I'm interested in learning your thoughts and views about the program and what the evaluation can do for you. What are some of your thoughts about the program?" (Notice "thoughts" is a vague, but neutral word that can prompt many different responses.)

We find it useful to begin in this rather general way. What stakeholders choose to tell the evaluator reflects their individual priorities. Making the initial questions more focused can result in missing important information or concerns. We start with a very open-ended question and are genuinely interested in what they may choose to say first. But evaluators will use additional questions and prompts to learn what stakeholders know about the program, what they think it is designed to achieve, what its strengths are, and what their concerns are. Some probing may enable the evaluator to learn the stakeholders' perceptions of the model or theory of the program while avoiding technical jargon. For example, if a theory-based approach seems appropriate, the evaluator might ask questions like: "What do you see as the major changes that will occur in students or clients as a result of participating in this program?" Then, "How do you think the program activities lead to those outcomes?" Or, "Which activities do you see as most important in leading to these goals?"

Having learned of the stakeholders' perceptions about the program, evaluators then move to learning more about what questions they want the evaluation to answer. There is no more important, or more frequently neglected, step for assuring that the evaluation will be used by its stakeholders. Evaluators can begin by asking: "What do you hope you will learn from the evaluation?" Or, "If I could collect information to answer any question about the program that you would like this evaluation to answer, what question would that be?" Like pickles in a jar,

evaluative questions are easier to get out after the first one has been extracted. Some probing may help stakeholders focus their thinking, using questions such as "What information would be most helpful to you to better manage or deliver the program? To decide whether to continue your support? Your participation in it?" Or for formative purposes: "Which program components or activities don't work as you thought they would? What are your concerns about them?"

If stakeholders overlook areas that others, or research, have suggested are important, the evaluator might ask, "Are you interested in X (fill in the area)?" "X" might be a specific program area (Are you interested in knowing more about how students first react to the new math approach?) or an evaluation stage (Are you interested in having a good description of whether the curriculum is being delivered as planned?) The question "What else would you like to know?" often produces abundant responses. This is no time to be judgmental or to point out that some suggested questions may currently be unanswerable. This is the time for generating all the evaluation questions possible. Weighing and selecting the subset of questions to be ultimately pursued will be done later, in the convergent stage. Evaluators should, however, briefly describe the process to all stakeholders interviewed so that they recognize that the questions will later be winnowed.

Figure 13.1 illustrates a possible sequence of questions in a stakeholder interview, leading from general questions intended to identify stakeholder views of the program, to more focused questions to identify their major evaluation questions. Additional specific procedures for guiding evaluator–participant interactions can be found in the writings of advocates of responsive and participative evaluation and

FIGURE 13.1 ***Information to Be Obtained in Interviews with Stakeholders***

1. What is your general perception of the program? What do you think of it? (Do you think well of it? Badly of it? What do you like about it? What do you not like? Why?)
2. What do you perceive as the purposes (goals, objectives) or guiding philosophy of the program? (Do you agree with these purposes or philosophy? Do you think the problems the program addresses are severe? Important?)
3. What do you think the theory or model for the program is? (Why/how do you think it works? How is it supposed to work? Why would the program actions lead to success on the program's objectives or criteria? Which program components are most critical to success?)
4. What concerns do you have about the program? About its outcomes? Its operations? Other issues?
5. What do you hope to learn from the evaluation? Why are these issues important to you?
6. How could you use the information provided by the answers to these questions? (Would you use it to make decisions, to enhance your understanding?)
7. What do you think the answer to the question is? (Do you already know it? Would you be concerned if the answer were otherwise?)
8. Are there other stakeholders who would be interested in this question? Who are they? What is their interest?

others (e.g., Abma & Stake, 2001; Cousins & Shula, 2008; Greene, 1987, 1988; King, 1998). Patton's (2008a) utilization-focused evaluation (UFE) provides additional guidance for the evaluator to learn about the information needs of stakeholders. By grounding the evaluation plan in the concerns of key people, the evaluator takes steps to ensure that the evaluation will be useful and responsive to constituents who may have differing points of view. For example, consider a leadership training program funded by an external foundation. Interviews with stakeholders of such a program might produce the following questions:

1. (From the program administrator) Are we running on time and within our budget? Are we meeting foundation expectations for this program? Is the program being implemented as planned? What changes have occurred and why? Are participants gaining the intended leadership skills at the desired level?

2. (From program staff) Are we delivering the program as planned? What changes are being made from the program model and why? How are trainees reacting to the program? Which sessions/methods work best? Which are worst?

3. (From participants toward whom the program is aimed) Have the leadership skills of participants really improved? Are they using them on the job? How? What portions of the program are most useful to participants?

4. (From the top managers in the organization) What evidence is there that the program is achieving its goals? Is this program having the desired impact on the units in which the trainees work? Would this program serve as a model for other change efforts in our organization? How is the work in this program changing our organization? What continuing expenses are going to exist once foundation support terminates?

5. (From the foundation) Is the program doing what it promised? What evidence is there that variables targeted for change have actually changed? How cost-effective is this program? Could the program be established in other settings? What evidence is there that the program will continue once foundation funds are terminated?

Using Evaluation Approaches as Heuristics

In exploring different approaches to evaluation in Part Two of this book, we noted that the specific conceptual frameworks and models developed under each approach play an important role in generating evaluation questions. This is one place in the evaluation process where the conceptual work done by the different evaluation theorists pays considerable dividends.

In reviewing the evaluation literature summarized in Part Two, the evaluator is directed toward certain questions. Sometimes a framework fits poorly and should be set aside, but usually something of value is suggested by each approach, as the following examples illustrate.

The program-oriented approach encourages us to use characteristics of the program as our guide to the evaluation. Objectives-oriented approaches lead us to

measure whether goals and objectives are achieved, but we may not evaluate all the objectives. Which ones hold the most interest for stakeholders? Why? Why are others not of interest? What standards will we use to determine if particular objectives are achieved? (How good must a performance be to be considered a success?) Theory-based approaches encourage us to learn about the program through developing or articulating the theory for the program, linking problems existing before the program to program actions and, then, to program outcomes. Evaluation questions may be based on any of the concepts that emerge in this process. This can include learning more about the problem the program is designed to address, investigating the critical program activities and their links to immediate outputs, or exploring or documenting immediate or long-term program outcomes.

Decision-oriented approaches lead evaluators to focus on information needs and decisions to be made. The particular management-oriented approach developed by Stufflebeam generates questions that typically arise at various stages of a program: the context (need), input (design), process (implementation), and product (outcomes) stages. Utilization-focused evaluation prompts evaluators to identify primary intended users and involve them heavily at this stage in identifying evaluation questions that meet their information needs (Patton, 2008a).

The participant-oriented approaches remind us that we should be sure to consider all stakeholders and should listen to what each group and individual has to say even during informal conversations. Differences in the practical participatory approaches and transformative participatory approaches remind us to consider what we hope to achieve through stakeholder participation. Practical participatory approaches are intended to increase use, often by involving a few stakeholders (managers or program staff) in more depth. Transformative participatory approaches are designed to empower stakeholders (program managers, staff, or clients) to evaluate their own programs and make effective decisions or to learn more about power differences in their setting or context and change those power arrangements. So, at this stage, these two approaches might take quite different directions. In most practical approaches, the evaluator is working in partnership with a few stakeholders to identify the evaluation questions, although other stakeholders are likely to be considered and interviewed. In transformative approaches and in evaluation capacity building (ECB), the evaluator takes more of a back seat to allow the stakeholders to learn through specifying and clarifying evaluation questions.

The consumer-oriented approach has generated many checklists and sets of criteria that may be of considerable value when deciding what components or characteristics to study in an evaluation or what standards to apply. The expertise-oriented approach has produced standards and critiques that reflect the criteria and values used by contemporary experts in education, mental health, social services, criminal justice, and other fields.

To the extent that these different approaches can stimulate questions that might not emerge elsewhere, they are important sources for evaluators to consider in the divergent phase of focusing the evaluation. As noted, many stakeholders are not familiar with the variety of issues an evaluation can address. We have found that

stakeholders will sometimes focus only on outcomes, assuming that an evaluation must measure outcomes. This is especially true in today's outcome-dominated culture. While in many cases such a focus is appropriate, often other concerns are more paramount, given the stage of the program and the needs of stakeholders. Posavac (1994) describes a case in which stakeholders' limited understanding of evaluation led to their advocacy for a summative evaluation when a formative evaluation was a more appropriate strategy. He argues that evaluators must "take an active role in helping their clients to understand what they really need" (p. 75). Similarly, Fitzpatrick (1989, 1992) describes how she persuaded stakeholders to allow her to examine the program activities in order to use them to interpret outcomes. State officials who oversaw the program were sure it was delivered as planned in spite of the fact that they rarely saw the program in operation. Her results provided important and surprising information that also helped them to interpret the program's successes and failures. Evaluation models can help the evaluator consider other areas of focus for the evaluation and can educate the stakeholder as to the myriad issues that evaluation can investigate.

Using Research and Evaluation Work in the Program Field

Many evaluators focus their work in a limited number of content areas or fields. Some evaluators work entirely in the field of education; others in such areas as mental health, health education, criminal justice, social services, training, or nonprofit management. In any case, the evaluator should be knowledgeable about theories and research findings in the area of the program and consider their relevance to the present evaluation. Even when the evaluator has worked often in the field of the program, in addition to interviewing stakeholders he should conduct a review of literature in the specific area of the program to learn more about what interventions have been found to be effective with what types of students or clients.

Existing research studies and theories can help the evaluator to develop or modify the program theory of stakeholders and suggest questions to guide the evaluation. Chen (1990), Weiss (1995, 1997), and Donaldson (2007) describe ways of using existing theory and research to develop or modify program models to guide the evaluation. Existing research and theory can be used to identify causes of the problem that the program is designed to address; to learn more about the specific factors that have succeeded, or failed, to remedy these problems; and to examine conditions that can enhance or impede program success with specific kinds of students or clients. The research literature can help the evaluator consider the likelihood that the program to be evaluated can succeed. It can be useful for the program evaluator to compare models in the research literature with the existing normative program model—the program developers' or stakeholders' model. Discrepancies between these models can suggest important areas for evaluation questions. For example, a program's stakeholders may argue that reading more books that are chosen to match individual students' interests and reading level will result in learning gains. Research may show that, for the particular reading problems

identified in this school, more directed teacher intervention is needed. Such differences between normative models and research do not, of themselves, prove the program will fail, but they do suggest that the evaluation should examine questions related to these differences. Published evaluations of similar programs can suggest not only questions to be examined but also methods, measures, and designs that might be productive for the evaluation study.

Familiarity with research, and confidence in one's consultative role, can help the evaluator in educating stakeholders about potential strengths and weaknesses of their approach. For example, enthusiastic sponsors of a merit pay program for teachers may need to be reminded of the mixed empirical results on merit pay in many other settings. (See Milanowski [2008]; Perry, Engbers, and Jun [2009].) Advocates for physical fitness and healthy eating programs in the schools may have neglected to consider the culture of the immigrant neighborhood and its impact on program success. The evaluator has a responsibility to raise such issues.

Commissions and task forces are sometimes formed by national, regional, or local governments to study problems of interest to governmental leaders. Such reports raise provocative questions and, although they occasionally make unsubstantiated claims, they usually reflect current social concerns, issues, and beliefs in the area or region. They may also serve to draw an informed evaluator's attention to issues that should be raised during a particular evaluation. Questions about important current issues may be omitted if the evaluator fails to raise them, with the result that the evaluation may be considered informative but devoid of information on the real issues facing the field today. Obviously, we are not proposing a faddish bandwagon approach to determine what questions will be addressed by an evaluation study, but it would be naïve indeed not to be aware of and consider the relevance of educational and social issues permeating current professional literature and other media.

Using Professional Standards, Checklists, Guidelines, and Criteria Developed or Used Elsewhere

In many fields, standards for practice have been developed. Such standards can often be useful either in helping to generate questions or in specifying criteria. (See, for example, *Standards of Evidence* developed by the Society for Prevention Research used for evaluations of prevention programs in the health arena. http://www.preventionreserach.org/StandardsofEvidencebook.pdf.) As with existing research and evaluation, standards can signal areas that may have been overlooked in focusing on the existing program. They are important resources for evaluators to have in their tool kits. If one were evaluating the success of an evaluation itself, the *Program Evaluation Standards* would obviously serve as an important guide to that evaluation. (We will cover metaevaluation, or evaluating evaluations, in Chapter 14.) Similarly, accrediting associations develop standards for judging institutions of higher education and hospitals. A review of their standards may prompt the evaluator and stakeholders to consider evaluation questions that had previously been neglected.

Asking Expert Consultants to Specify Questions or Criteria

Evaluators are often asked to evaluate programs outside their areas of content expertise. For example, an evaluator may be called on to evaluate a school's reading program, even though he or she knows little about reading programs. The stakeholders can provide valuable expertise in helping orient the evaluator to the particulars of the program. But, in some cases, the evaluator may also want to make use of a consultant with expertise in the content of the program to provide a more neutral and broader view than the evaluator can gain from program staff. Such consultants can be helpful in suggesting evaluation questions and criteria that reflect current knowledge and practice.

In the case of evaluating a school reading program, for example, the consultant could be asked not only to generate a list of evaluation questions to be addressed, but also to identify previous evaluations of reading programs, standards set by professional organizations such as the International Reading Association, and research on the criteria and methods for evaluating reading programs. If there is concern about possible ideological bias, the evaluator might employ more than one independent consultant.

Using the Evaluator's Professional Judgment

Evaluators should not overlook their own knowledge and experience when generating potential questions and criteria. Experienced evaluators are accustomed to describing the object of the evaluation in detail and looking at needs, program activities, and consequences. Perhaps the evaluator has done a similar evaluation in another setting and knows from experience what questions proved most useful. Professional colleagues in evaluation and the content field of the program can suggest additional questions or criteria.

Evaluators are trained, at least in part, to be skeptics—to raise insightful (one hopes) questions that otherwise might not be considered. This training is never more valuable than during the divergent phase of identifying evaluation questions and criteria, because some important questions may be omitted unless the evaluator raises them. House and Howe (1999), though advocating deliberative democracy in evaluation to give a voice to less powerful stakeholders, make it quite clear that, in their view, evaluators have the authority and responsibility to make use of their own expertise. Certainly, evaluators bring in, and balance, the values and views of different stakeholders. Yet, evaluators play a key role in leading the evaluation and, therefore, must make use of their expertise in knowing the types of evaluation questions that can be most usefully addressed at different program stages within the resources and constraints of the particular evaluation context.

Even in the context of an empowerment evaluation, evaluators recognize the importance of the evaluator's role at that stage. In their description of an empowerment evaluation, Schnoes, Murphy-Berman, and Chambers (2000) make it clear that they work hard to empower their users, but, ultimately, they write,

"Whose standards of accountability should prevail in defining project outcomes, particularly if the clients' understanding and notions of what consists of valid measurable results are at variance with the standards endorsed by evaluators?" (p. 61). Evaluators are hired for their knowledge and expertise, and adding to the generation of evaluation questions based on that knowledge and experience is not only appropriate, but mandatory in many situations.

Experienced and insightful evaluators looking at a new project might raise questions such as:

- Are the purposes the project is intended to serve really important? Is there sufficient evidence of need for the project as it is designed? Are other more critical needs going unattended?
- Are the goals, objectives, and project design consistent with documented needs? Are the program activities, content, and materials consistent with students' or clients' needs, goals, and objectives?
- Have alternative strategies been considered for accomplishing the project's goals and objectives?
- Does the program serve the public good? Does it serve democratic goals? Community goals?
- What are some of the unintended side effects that might emerge from this program?

Evaluators might ask themselves the following questions:

- Based on evaluations of other, similar projects, what questions should be incorporated into this evaluation?
- Based on my experience with other, similar projects, what new ideas, potential trouble spots, and expected outcomes or side effects can be projected?
- What types of evidence will be accepted by different stakeholders? Can their criteria for evidence be successfully addressed by the current evaluation questions?
- What critical elements and events should be examined and observed as the project develops?

Summarizing Suggestions from Multiple Sources

Somewhere in the divergent process the evaluator will reach a point of diminishing returns when no new questions are being generated. Assuming that each available resource has been tapped, evaluators should stop and examine what they have obtained: usually, long lists of several dozen potential evaluation questions, along with potential criteria. So that the information can be more readily assimilated and used later, the evaluator will want to organize the evaluation questions into categories. Here certain evaluation frameworks or approaches, such as Stufflebeam's (1971) CIPP model, empowerment evaluation's ten-step approach (Wandersman et al., 2000), or Rossi's program theory (see Chapter 12), may be useful. The evaluator might adopt labels from one of these frameworks or

create a new set of categories tailored to the study. Regardless of the source, having a manageable number of categories is essential in organizing potential questions and communicating them to others. Here is a sample of possible questions that might arise in the divergent phase for planning an evaluation of a conflict-resolution program in the schools:

Needs Assessment or Context

1. What kinds of conflict occur among students in the schools? Who is most likely to be involved in a conflict (age, gender, characteristics)? What is the nature of the conflict?
2. How were conflicts resolved before the program? What kinds of problems occurred as a result of this strategy?
3. What communication skills do the students have that conflict resolution could build on? What problems do the students have that might hinder the learning or use of conflict-resolution skills?
4. How many conflicts currently occur? How frequent is each type?
5. What effects do the current conflicts have on the learning environment? On the management of the school? On the motivation and abilities of the students? On the retention of good teachers?

Process or Monitoring

1. Are the conflict-resolution trainers sufficiently competent to provide the training? Have the appropriate personnel been selected to conduct the training? Should others be used?
2. Do the students selected for training meet the specified criteria for the target audience?
3. What proportion of students participates in the complete training program? What do these students miss by participating in the training (opportunity costs)?
4. Does the training cover the designated objectives? Is the training at the necessary level of intensity or duration?
5. Do students participate in the training in the intended manner?
6. Where does the training take place? Is the physical environment for the training conducive to learning?
7. Do other teachers in the school encourage use of the conflict-resolution strategies? How? Do the teachers use these strategies themselves? How? What other strategies do they use?

Outcomes

1. Do the students who have received the training gain the desired skills? Do they believe the skills will be useful?
2. Do the students retain these skills one month after the completion of training?
3. What proportion of the students has used the conflict-resolution strategies one month after program completion? For those who have not used the strategies, why not? (Were they not faced with a conflict, or were they faced with a conflict but used some other strategy?)

4. Under what circumstances were students most likely to use the strategies? Under what circumstances were they least likely to use them?
5. How did other students support or hinder the students' use of the strategies?
6. Did the students discuss/teach the strategies to any others?
7. Was the incidence of conflicts reduced at the school? Was the reduction due to the use of the strategies?
8. Should other students be trained in the strategy? What other types of students are most likely to benefit?

It will be obvious to thoughtful evaluators and stakeholders that it is not feasible or even desirable to address all these questions in any one study. Practical considerations must limit the study to what is manageable. Some questions might be saved for another study; others might be discarded as inconsequential. Such winnowing is the function of the convergent phase.

Selecting the Questions, Criteria, and Issues to Be Addressed: The Convergent Phase

Cronbach (1982) gives a good introduction to the need for a convergent phase of evaluation planning:

> The preceding section [the divergent phase] spoke as if the ideal were to make the evaluation complete, but that cannot be done. There are at least three reasons for reducing the range of variables treated systematically in an evaluation. First, there will always be a budget limit. Second, as a study becomes increasingly complicated, it becomes harder and harder to manage. The mass of information becomes too great for the evaluator to digest, and much is lost from sight. Third, and possibly most important, the attention span of the audience is limited. Very few persons want to know all there is to know about a program. Administrators, legislators, and opinion leaders listen on the run.
>
> The divergent phase identifies what could possibly be worth investigating. Here the investigator aims for maximum bandwidth. In the convergent phase, on the contrary, he decides what incompleteness is most acceptable. He reduces bandwidth by culling the list of possibilities. (p. 225)

No evaluation can answer responsibly all the questions generated during a thorough, divergent planning phase. The program is at a given stage, so some questions are appropriate for that stage and others are not. Similarly, the budget, timeframe, and context will restrict the questions that can be addressed. So the question is not whether to winnow these questions into a manageable subset, but who should do it and how.

Who Should Be Involved in the Convergent Phase?

Some evaluators write and behave as if selecting crucial, practical evaluation questions were the sole province of the evaluator. Not so. In fact, under no circumstances should the evaluator assume sole responsibility for selecting the

questions to be addressed or the evaluative criteria to be applied. This task requires close interaction with stakeholders. The sponsor of the evaluation, key audiences, and individuals or groups who will be affected by the evaluation should all have a voice. Often, this is the time to establish an advisory group with representatives of different stakeholder groups to winnow the evaluation questions and to serve as a sounding board and advisor for the remainder of the evaluation.

Indeed, some evaluators are content to leave the final selection of questions to the evaluation sponsor or client. Certainly this lightens the evaluator's task. In our view, however, taking that easy course is a disservice to the client. Lacking the advantage of the evaluator's special training and experience, the client may well wind up posing a number of unanswerable or costly questions for the study.

How Should the Convergent Phase Be Carried Out?

How can the evaluator work with the multiple stakeholders to select the questions for the evaluation? To begin with, the evaluator can propose some criteria to be used to rank the potential evaluation questions. Cronbach and others (1980) suggest the following criteria:

> So far we have encouraged the evaluator to scan widely; only in passing did we acknowledge that all lines of inquiry are not equally important. How to cut the list of questions down to size is the obvious next topic.
>
> . . . simultaneous consideration is given to the criteria . . . [of] prior uncertainty, information yield, costs, and leverage (that is, political importance). These criteria are further explained as follows: The more a study reduces uncertainty, the greater the information yield and, hence, the more useful the research.
>
> Leverage refers to the probability that the information—if believed—will change the course of events. (pp. 261, 265)

We draw on Cronbach's thinking in proposing the following criteria for determining which proposed evaluation questions should be investigated:

1. *Who would use the information? Who wants to know? Who will be upset if this evaluation question is dropped?* If limitless resources were available, one could argue that (except for invading rights of privacy) anyone who wishes to know has, in a democratic society, the right to information about what is evaluated. Rarely are resources limitless, however, and even if they were, prudence suggests a point of diminishing returns in collecting evaluative information. Therefore, if no critical audience will suffer from the evaluator's failure to address a particular question, one might well give it a lower ranking or delete it. What is a critical audience? That audience will vary with the context of the evaluation. In some cases, critical audiences are decision makers because a decision is imminent and they are uninformed. In other cases, previously uninvolved or uninformed stakeholders (program participants, family members of participants, emerging interest groups) are the critical audiences by virtue of their previous lack of involvement and their need to know.

2. *Would an answer to the question reduce present uncertainty or provide information not now readily available?* If not, there seems little point in pursuing it. If the answer already exists (or the client who would use the information thinks he or she knows the answer), then the evaluation should turn to other questions for which answers are not available.

3. *Would the answer to the question yield important information? Would it have an impact on the course of events?* Some answers satisfy curiosity but little more; we call them "nice to know" questions. Important questions are those that provide information that might inform action on substantive issues concerning the program and its clients. They may address areas considered problematic by stakeholders with the motivation and means to make or influence changes.

4. *Is this question merely of passing interest to someone, or does it focus on critical dimensions of continued interest?* Priority should be given to critical questions of continuing importance. Program theory can help identify the critical dimensions of the program that the evaluation might address.

5. *Would the scope or comprehensiveness of the evaluation be seriously limited if this question were dropped?* If so, it should be retained, if possible. In some cases, however, comprehensiveness, evaluating every aspect of the program, is less important than evaluating certain areas of uncertainty in depth. The evaluator and stakeholder should consciously consider the issues of breadth versus depth in their selection of evaluation questions.

6. *Is it feasible to answer this question, given available financial and human resources, time, methods, and technology?* Limited resources render many important questions unanswerable. Better to delete them early than to breed frustration by pursuing impossible dreams.

The six criteria just noted can be cast into a simple matrix (see Figure 13.2) to help the evaluator and client narrow the original list of questions into a manageable subset. Figure 13.2 is proposed only as a general guide and may be adapted or used flexibly. For example, one might expand the matrix to list as many questions as exist on the original list, then simply complete the column entries by answering yes or no to each question. Alternatively, questions could be assigned a numerical rating, which offers the advantage of helping to weight or rank questions.

Working with Stakeholders. However the matrix is used, the evaluator and client (or the advisory group or representatives of other stakeholders if no advisory group exists) should work together to complete it. Although the evaluator may have the say on what is feasible, the relative importance of the questions will be determined by the client and other stakeholders. Scanning the completed matrix quickly reveals which questions are not feasible to answer, which are unimportant, and which can and should be pursued.

In some cases, the evaluator may take a leadership role in sorting the questions, combining some that could be addressed with the same methods, editing the

FIGURE 13.2 ***Matrix for Ranking or Selecting Evaluation Questions***

	Evaluation Question						
Would the evaluation question . . .	*1*	*2*	*3*	*4*	*5*	*. . .*	*n*
1. Be of interest to key audiences?							
2. Reduce present uncertainty?							
3. Yield important information?							
4. Be of continuing (not fleeting) interest?							
5. Be critical to the study's scope and comprehensiveness?							
6. Have an impact on the course of events?							
7. Be answerable in terms of							
A. Financial and human resources?							
B. Time?							
C. Available methods and technology?							

questions for clarity, considering the feasibility of each and the criteria, and developing a list for review by the client or group. The list might be organized into groups: recommended questions, potential ones given sufficient interest or more resources, and those which might be dropped at this point.

The sponsor or client will likely want to add or subtract selected questions, possibly negotiating an increased or reduced scope for the study or debating the rationales for adding or dropping certain questions. In these important negotiations, evaluators may find it necessary to defend their own professional judgment or the interests of unrepresented stakeholders. This can be difficult. If the sponsor or client demands too much control over the selection of evaluation questions (e.g., requiring inclusion of unanswerable questions or those likely to yield one-sided answers, or denying the needs of certain stakeholder groups), the evaluator must judge whether the evaluation will be compromised. If it is, it may be in the best interest of all concerned to terminate the evaluation at this point, though certainly the evaluator should take this opportunity to educate the sponsor on what evaluation can do and the ethical guidelines of evaluation practice. Conversely, evaluators must refrain from insisting on their own preferred questions and overriding legitimate concerns of the sponsor or client.

Usually the evaluator and client can agree on which questions should be addressed. Reaching a congenial consensus (or compromise) goes far toward establishing the sort of rapport that turns an evaluation effort into a partnership in which the client is pleased to cooperate. A feeling of shared ownership greatly enhances the probability that the evaluation findings to come will ultimately be used.

If the evaluator and the client select the final evaluation questions, the evaluator has an obligation to inform other stakeholders of the focus of the evaluation and the final questions. To facilitate this dialogue, the evaluator can provide a list of questions to be addressed with a short explanation indicating why each is important. If the matrix (Figure 13.2) is used, a copy can be provided. The list of questions and/or

matrix should be shared with all important stakeholders in the evaluation. They should be told that this tentative list of questions is being given to them for two reasons: (1) to keep them informed about the evaluation, and (2) to elicit their reactions, especially if they feel strongly about adding or deleting questions. Sufficient time should be set aside for their review before the final list is produced.

Concerned comments merit a direct response. The evaluator should meet with any stakeholders who are dissatisfied with the list of questions and with the sponsor, if need be, to discuss and resolve concerns to everyone's satisfaction before continuing. To push for premature closure on legitimate issues surrounding the scope of the evaluation is one of the worst mistakes the evaluator can make. Unresolved conflicts will not go away, and they can be the undoing of an otherwise well-planned evaluation.

One caution: A timeworn but effective ploy used by those who wish to scuttle an unwanted evaluation is to raise unresolvable objections. The astute evaluator should recognize strident insistence on including biased or unanswerable questions. This is where an advisory committee of stakeholders, including the conflicting parties, can be particularly useful. The committee can be given the task of hearing and making recommendations on the evaluative questions to be addressed. Other stakeholders and the evaluator can then work to clarify the objections, modify the questions as appropriate, and move toward consensus.

Specifying the Evaluation Criteria and Standards

A final step at this stage is specifying the criteria that will be used to judge the program and the standards for success. This step comes quite naturally after consensus is achieved on the final evaluation questions. Some questions, because of their very nature, require the specification of criteria and the standards for those criteria for the evaluator to make the judgment the question requires. In most cases, these are evaluation questions that will require judgments of success or failure to answer the question. The criteria and standards are to assist in making those judgments.

First, let us talk a bit about criteria and standards, because they are occasionally confused. Criteria are the factors that are considered important to judge something. Jane Davidson (2005) defines criteria as "the aspects, qualities, or dimensions that distinguish a more meritorious or valuable evaluand [the object being evaluated] from one that is less meritorious or valuable" and goes on to comment that "[c]riteria are central to any evaluation" (p. 91). Standards, then, are the level of performance expected on each criterion. Standards are subsets of the criteria. In discussing how standards for evaluation might be set using participative or collaborative approaches, Cousins and Shula echo Davidson's views:

> Program evaluation requires judgment about program merit and/or worth and/or significance, and judgment requires comparison between what is observed, through the systematic gathering of data, and some criterion or set of criteria. In evaluating program quality the question naturally arises, 'How good is

> good enough?' This question not only is central to evaluation, but also, we would argue, it is the essential characteristic differentiating evaluation from social sciences research. To answer the question, a program standard must be set. (2008, p. 139)

If the standard is achieved, the program can be judged successful on that criterion. An example may clarify: the criteria one might use to judge the quality of a training program could include attendance by the intended participants, participants' satisfaction with the program, their learning of key concepts, and application of those concepts on the job. A standard for attendance at a training program might be that 95% of the employees who were designated to attend the training completed the program. A standard for application on the job might be that 75% of those completing the program used at least one of the three curriculum planning strategies presented at the program in planning curricula with their team of teachers in the next month. And how are the numbers in these standards determined? Through dialogue with stakeholders, the evaluators' expertise, and information from other programs, research, and evaluations. But achieving consensus is not an easy task. Let us describe the process.

It is best to identify criteria and the standards for them before moving further in the evaluation, certainly before data collection begins. New criteria may emerge as the evaluators learn more about the program and stakeholders learn more about evaluation. However, stakeholders will have referred to or discussed criteria, directly or indirectly, in their conversations with the evaluator and other stakeholders during the divergent and convergent phases. So, now is the time to seek clarity on some of those criteria and to consider standards. It is important for the evaluator and different stakeholder groups to have some idea of the levels of program performance that are considered acceptable now, rather than waiting until data are analyzed. Achieving such consensus can prevent later disagreements. Without a consensus on expected levels of performance, program advocates can claim that the obtained level of performance was exactly the one desired, and program detractors can claim that the same level of performance is insufficient for program success. Further, agreeing on standards prior to obtaining results can be very useful in helping groups to be clear and realistic about their expectations for program success. In today's political environment, the goals for programs are often unrealistically high or too vague and surrounded by language that implies astonishing rates of success—100% graduation rates, 100% literacy, 100% employment! It is better to have high expectations than low ones; however, it is necessary to have a realistic discussion of what can be accomplished, given the program activities and their frequency or intensity (sometimes referred to as "dosage") and the qualifications and skills of those delivering the program, so that all have a similar understanding of what success (or failure) looks like.

Criteria will emerge from the same sources as the evaluation questions. In other words, conversations and meetings with stakeholders, research and evaluation found through reviews of literature, professional standards, experts, and the evaluators' own judgment can all suggest factors that should be used to judge the success of a program or one of its components. For example, the criteria to judge a training

program are well known in training evaluation models and literature (reaction, learning, behavior, results) (Kirkpatrick, 1983; Kirkpatrick & Kirkpatrick, 2006).

The specification of standards of performance can be a complex area fraught with uncertainty. Stakeholders can be reluctant to specify numbers that reflect success because they genuinely do not know what to expect. In some cases, the program is too new or the specification of standards is too divisive to create realistic and valid standards. The evaluator should be sensitive to the fact that staff, when pushed into developing standards, may feel defensive and develop standards that they feel certain to achieve. Such standards may not reflect others' goals for the program. Program opponents or policymakers advocating for resources may suggest standards that are unlikely to be achieved. Nevertheless, in many cases, stakeholders' discussions of expectations, moving to a development of standards, can ultimately be quite useful.

Where similar programs exist, they can be examined for standards of performance. The research literature can be very helpful here. In almost every case, there are programs that have attempted to bring about similar changes. A search of the literature can find a variety of reading programs delivered to similar students, substance abuse treatment programs delivered to similar clients, or job training programs for similar populations. Evaluators should examine this literature to learn more about the amount of change such programs have been able to achieve. To stimulate discussion, the evaluator might present stakeholders with a list of ten other programs, cited as successes in the literature, and the amount of change (perhaps measured as a proportion of change from the pre-test or an effect size) that was achieved in each. Such data can be sobering: tough problems are not solved immediately. But, such data provide boundaries for developing standards. Do stakeholders see their program, their resources, and their clients as similar to these? Much better? Much worse? How should the successes (or failures) found with other programs be adjusted to reflect the expected performance of the program to be evaluated?

Standards can be absolute or relative. We will briefly discuss each type here. Evaluators might first work with stakeholders to consider which type of standard will be most useful for their program in the political and administrative context in which it operates.

Absolute Standards

Sometimes policy will require the specification of an absolute standard. In the United States today, No Child Left Behind (NCLB) requires states to develop specific standards to assess students' educational progress. These standards are absolute, not relative. That is, they reflect an amount of knowledge expected of students at various grade levels. Typically, there are multiple levels. If these standards overlap with the goals of the program, as will sometimes be the case in educational settings, the standards may be appropriate to use in the evaluation. (This would only be the case in circumstances when the standards are realistic for the student population. In some cases, state standards have been established

primarily for political purposes and do not necessarily reflect realistic, or feasible, outcomes. See the American Evaluation Association's statement [2002] on high-stakes testing at www.eval.org/hstlinks.htm.) Similarly, accreditation requirements or standards for care of patients can suggest absolute standards that may be used.

When absolute standards do not exist, the evaluator can begin with a review of literature seeking research or evaluation results on similar programs, as described earlier, and use those results as a stimulus for stakeholder discussion. Or if stakeholders seem ready, evaluators can seek input from knowledgeable stakeholders about their expectations on each criterion. As evaluators learn the range of expectations, they can then lead a discussion of proposed standards with the key stakeholders or an advisory group. Thus, if attendance were an important criterion for success, the evaluator might ask, "What proportion of students do you expect to complete the program? 100 percent? 90 percent? 75 percent?" or "How much of a reduction in disciplinary incidents do you expect to occur as a result of the program? 75 percent? 50 percent? 25 percent?" Such questions can prompt a frank discussion of expectations that will be invaluable in judging program results. Evaluators should avoid having program staff purposely set standards too low (to ensure program success) or having program opponents set them too high (to guarantee failure). Working with a group of stakeholders with different perspectives can help avoid such a situation.

Relative Standards

Some argue that absolute standards such as those discussed in the previous section are unnecessary when the study will involve comparisons with other groups. Thus, Light (1983) argues that outcomes superior to those achieved with a placebo control or comparison group are sufficient to demonstrate program success. Such relative standards are certainly the standards used in other fields such as medicine and pharmacology; the common standard is whether the new drug or procedure results in better cure rates or fewer side effects than the currently used drug or procedure. No absolute standard is set. Scriven (1980, 1991c, 2007) advocates comparing programs with available alternatives, in other words, the programs or policies that students or clients will receive instead of this program. Such comparisons reflect the actual choices that policy makers and others will be making. The standard in such cases may be that the new program is significantly better than the alternative method. Other relative standards compare program performance with past performance. Thus, the three-step process in empowerment evaluation first develops a baseline measure of program performance and then judges future success by comparison with that baseline (Fetterman & Wandersman, 2005). In fact, such comparisons with past performance are the comparisons frequently made in the private sector (sales now, compared with the same period last year) and in economic policy (unemployment rates now, compared with last month or this time last year).

Remaining Flexible during the Evaluation: Allowing New Questions, Criteria, and Standards to Emerge

Evaluations can be flawed by evaluators who relentlessly insist on answering the original questions, regardless of intervening events, changes in the object of the evaluation, or new discoveries. During the course of an evaluation, many occurrences—for example, changes in scheduling, personnel, and funding; unanticipated problems in program implementation; evaluation procedures that are found not to work; lines of inquiry that prove to be dead ends; new critical issues that emerge—require new or revised evaluation questions. Because such changes cannot be foreseen, Cronbach and his associates (1980) propose that:

> Choice of questions and procedures, then, should be tentative. Budgetary plans should not commit every hour and every dollar to . . . the initial plan. Quite a bit of time and money should be held in reserve. (p. 229)

When changes in the context or object of the evaluation occur, the evaluator must ask whether that change should affect the list of evaluation questions. Does it make some questions moot? Raise new ones? Require revisions? Would changing questions or focus in the middle of the evaluation be fair? The evaluator should discuss any changes and their impact on the evaluation with the sponsor, client, and other stakeholders. Allowing questions and issues to evolve, rather than committing to an evaluation carved in stone, fulfills Stake's (1975b) concept of responsive evaluation discussed in Chapter 8. and such flexibility and responsiveness is advocated by today's participatory approaches (Cousins & Shula, 2008).

A word of warning, however: Evaluators must not lose track of questions or criteria that—despite possible changes—remain important. Resources should not be diverted from vital investigations just to explore interesting new directions. Flexibility is one thing, indecisiveness another.

Once the evaluation questions, criteria, and standards have been agreed upon, the evaluator can complete the evaluation plan. The next steps in the planning process are covered in Chapter 14.

Major Concepts and Theories

1. Evaluation questions give focus to the evaluation. They specify the information the evaluation will provide and guide choices for data collection, analysis, and interpretation.

2. The divergent phase of question development may be conducted with many stakeholders or with a few primary intended users with input from other stakeholders. It results in the development of a list of potential evaluation questions and concerns.

3. Other sources for questions include evaluation approaches, existing standards in the field, research literature, content experts, and the evaluator's own experience.

4. The convergent phase concerns winnowing down the questions for the final evaluation. Questions that are retained should have high, direct potential use to important and/or many stakeholders. Questions may be further culled based on the cost and feasibility of providing valid answers.

5. Criteria specify the factors that will be used to consider a program's success. Standards indicate the level of performance a program must reach on the criteria to be considered successful.

6. Criteria may be identified using the sources for evaluation questions.

7. Setting standards is a sensitive but important process. The dialogue involved in establishing standards can be useful in achieving clarity and consensus amongst stakeholders and in helping make later judgments on the program. Standards may be relative or absolute.

Discussion Questions

1. Do you think it is better to involve many stakeholders in the divergent phase or only a few but ones who are committed to using the evaluation, such as primary intended users? Discuss your reasoning.

2. As we discussed, evaluators and stakeholders bring different types of expertise to identifying evaluation questions and developing the evaluation plan. Contrast the knowledge evaluators bring to the table with that of stakeholders. Consider particular stakeholder groups and how their expertise can best be used in the evaluation.

3. Which of the evaluation approaches discussed in Part Two do you think are most useful in helping develop evaluation questions?

4. As an evaluator, what would you do if a client or other stakeholder were to push adamantly for a biased or unanswerable question?

5. Discuss the potential advantages and disadvantages of using absolute versus relative standards to judge the success of a program.

Application Exercises

1. Consider an evaluation that would have meaning for you and your organization or employer. (If you are not employed or do not know a program well enough to consider it, you might select your graduate program or try some recent, highly publicized program or policy being considered by city or state officials.) Using what you now know about identifying evaluation questions during the divergent phase, generate a list of evaluation questions you would want to address. What steps might you take to identify other questions?

2. What method would you use to cull the above questions in the convergent phase, knowing your organization and the issues involved? Do the criteria in Figure 13.2 serve the purpose? Would you modify them? Which questions do you think should take priority? Why? Which should be dropped? Why?

3. Which of your questions would benefit from criteria and standards? If the evaluation is summative, do the questions convey all the important criteria for the program? Should other questions or criteria be added? Now, set standards for each question as appropriate. Discuss your rationale for each standard.
4. Interview two fellow students separately about evaluating your graduate program. Start broadly with their knowledge of the program, their perceptions of its strengths and weaknesses, and their concerns. Then, discuss questions each thinks the evaluation should address. What did you learn from the interviews? From conducting the process? What differences did you discover? Why do you think these differences exist? How different do you think the responses of other stakeholders (faculty, current or future employers of students, university administrators, advisors) might be? If possible, interview some of them.
5. Obtain a copy of a report from a completed evaluation study. (Each of the case studies recommended at the ends of chapters includes a reference to the evaluation report. Read the report and the interview.) Consider the questions that were addressed. Were there any critical oversights? Was the evaluation formative or summative? Was the focus on needs assessment, monitoring program activities, or examining outcomes? Were criteria and/or standards explicitly stated? If not, was their omission acceptable? Why or why not? If they were stated, on what grounds were they developed? Do you agree with the criteria? Would you have added others? Were standards set at the appropriate level?

Relevant Evaluation Standards

We consider the following evaluation standards, which are listed in their entirety in Appendix A, to be relevant to this chapter's content:

U2—Attention to Stakeholders
U3—Negotiated Purposes
U4—Explicit Values
U5—Relevant Information
U6—Meaningful Processes and Products
U8—Concerns for Consequences and Influence
F2—Practical Procedures
F3—Contextual Viability
P1—Responsive and Inclusive Orientation
P4—Clarity and Fairness
P6—Conflicts of Interest
A1—Justified Conclusions and Decisions
A2—Valid Information
A3—Reliable Information
E1—Evaluation Documentation

Case Studies

We recommend two interviews that illustrate issues discussed in this chapter: *Evaluation in Action,* Chapters 2 (James Riccio) and 6 (Debra Rog).

In Chapter 2, Jim Riccio discusses some of the questions of interest to the California state legislature in his evaluation of an early welfare reform program. He also discusses some of his own criteria and standards in responding to Fitzpatrick's questions concerning what constitutes success. Journal source: Fitzpatrick, J. L., &

Riccio, J. A. (1997). A dialogue about an award-winning evaluation of GAIN: A welfare-to-work program. *Evaluation Practice*, *18*, 241–252.

In Chapter 6, Debra Rog demonstrates her flexibility and discusses how the evaluation questions changed as the evaluation helped her and her client learn more about homeless families and some of the data did not match their expectations. She also discusses her criteria for judging program success. Journal source: Fitzpatrick, J. L., & Rog, D. J. (1999). The evaluation of the Homeless Families Program. A dialogue with Debra Rog, *American Journal of Evaluation*, *20*, 562–575.

Suggested Readings

Cousins, J. B., & Shula, L. M. (2008). Complexities in setting program standards in collaborative evaluation. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation,* pp. 139–158. New York: Guilford Press.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass.

Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications.* New York: Erlbaum Associates.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

14

Planning How to Conduct the Evaluation

Orienting Questions

1. What are some activities or functions common to all evaluations that must be considered in planning any evaluation study?
2. What should be specified in the evaluation plan?
3. What is the role of the client and other stakeholders in developing the plan?
4. How can you organize time, responsibilities, and resources so that all evaluation tasks are accomplished in a first-rate and timely manner?
5. What factors must be considered when developing evaluation budgets?
6. Why would a formal evaluation contract or agreement between evaluator and client be useful?
7. What is a metaevaluation and how can it be useful to an evaluation?

Much has been said in earlier chapters about the need to *focus* the evaluation study—to understand what is to be evaluated, why the evaluation has been proposed, what the evaluation's sponsor, client, and other stakeholders want to learn, and what criteria they would use to make judgments. But is this evaluation planning? Yes. When the focus of a study has become clear, is the evaluation plan complete? No, because focusing is only one part of developing an evaluation plan.

To explain the relationship between focusing and planning an evaluation, we turn to the work of Stufflebeam (1968, 1973b; Stufflebeam & Shinkfield, 2007). He proposes that one should first focus the evaluation to determine what information is needed. He also proposes four functions that are common to conducting any evaluation: information collection, organization, analysis, and reporting. To

develop an evaluation design, Stufflebeam maintains that one must plan how each of these functions would be carried out. Finally, he suggests that developing a plan for administering the evaluation is an integral part of an evaluation design. The resultant structure for developing evaluation designs includes these six activities/functions:

1. *Focusing* the evaluation
2. *Collecting* information
3. *Organizing* information
4. *Analyzing* information
5. *Reporting* information
6. *Administering* the evaluation (Stufflebeam & Shinkfield, 2007)

In Chapters 11 through 13 we dealt with various aspects of *focusing the evaluation* (phase 1 in the list). Understanding the origin and context of a proposed evaluation and identifying and selecting the evaluation questions, criteria, and standards most appropriate for the study are the major aspects of focusing the evaluation. In this chapter, we discuss how phases 2 through 6—collecting, organizing, analyzing, and reporting information and administering the evaluation—ought to be considered in the evaluation plan. (Chapters 15 through 17 will then provide more detail on carrying out these phases.) Before addressing these topics, we remind the reader of two important points:

1. *Evaluations should be conducted in a flexible manner.* One should not infer that the steps involved in evaluation are sequential and linear. We might almost as conveniently have used Stake's (1975b) "clock" here (see Figure 8.2), which emphasizes that one may move back and forth among evaluation functions, from data analysis to more data collection, to reporting, back to reanalysis, and so on.

2. *The evaluator should have a clear understanding of the purpose and role of the evaluation.* In earlier chapters, we outlined several different approaches to evaluation and described how each might be used to perform different evaluation roles. Then we provided practical guidelines (especially in Chapters 12 and 13) to help the evaluator focus the study. It would be difficult for an evaluator to go through activities such as the ones we have proposed without developing a fairly clear notion of the role the evaluation will play and a general idea of the type of evaluation study that will best suit that role.

Yet, far too often evaluators arrive at this point, after considerable interaction with the client and other stakeholders, and are still unable to articulate clearly the purposes or focus of the evaluation. By now, evaluators should understand the information needs of the stakeholders regarding the program. It should be clear whether the evaluation is primarily formative or summative and whether the focus is on needs assessment, describing program processes, or assessing program outputs or outcomes. The evaluator should also have a good sense for the program, its theory and intended impacts, the stage of the program, and the

context in which it operates. Now is the time to begin considering how the evaluation will be conducted to answer the identified evaluation questions. Evaluators should develop a plan that will convey the information to be collected, the methods and procedures to be used, and how the evaluation will be managed. In Part Four, we will discuss data collection in detail. Here, we will touch on the critical decisions evaluators need to make to finalize an evaluation plan that will convey to stakeholders and to the evaluation team how the evaluation will be conducted. The plan will begin with the evaluation questions, but will then specify appropriate designs and data collection strategies; procedures for collecting the data; and the means for analyzing, interpreting, and reporting the information. The management plan will describe the plan for staffing and managing the evaluation and its costs. Sections in this chapter will address the decisions to be made on each of these issues.

Developing the Evaluation Plan

In Chapter 13, we dealt with identifying and selecting the questions that the evaluation study should answer. Once the questions are known, the next logical step is to determine what information is needed to answer each question. For example, consider the monitoring question, "Have the critical program activities been delivered as planned?" To answer this question, the evaluator would need to know, among other things, which activities were identified as critical and the details of how they would be delivered, that is, to what kinds of students or clients, in what manner, with what intensity or duration, by staff with what types of skills and training.

The information needs in the preceding example may seem relatively straightforward, but in practice they are often much more complex. Many monitoring studies provide great detail on the wrong things. The program theory and logic models and the extensive communication with program developers and managers are necessary to specify the activities to be monitored and the critical characteristics to be described.

Consider another example in which the outcome question is, "What impact does the computer-based WANDAH program have on the writing performance of high school students in the Jefferson High WANDAH writing classes?" To answer the question it would be necessary to first choose the appropriate design. Is the program at a pilot stage, suggesting that the design should be descriptive? Is a summative decision to be made, making causal attributions more appropriate? What types of evidence are most acceptable to key stakeholders? How will causality be established? Another critical issue, of course, is how writing performance will be measured. Writing performance could be viewed either holistically or analytically, or both. Holistic measures of students' writing ability might involve judgments, made by panels, of the overall quality of student writing samples before and after exposure to WANDAH. An analytic approach might include measures of syntactic density, numbers of T-units, percent of *to be* verbs, or average sentence

length before and after using WANDAH. Students' writing performance might also be measured according to the extent and effectiveness of revisions from one draft to another. In this armchair example, we can avoid the choice, but were we actually conducting the study we would need to decide which type of design would be best and precisely what constructs and measures should be used to answer the question.

Obviously, the evaluator should involve the client and other stakeholders in deciding what information would best answer each evaluation question, but the evaluator also plays an active and pivotal role. He or she may have worked on evaluating writing programs in the past and be familiar with current, valid, and feasible means for assessing students' writing abilities. If not, literature searches of research and evaluations of writing programs, and consultations with writing experts, will suggest possible methods. The evaluator, as the methodologist, has the expertise to review methods discussed in the literature and to consider their appropriateness for the WANDAH program, its stakeholders, and its setting. In particular, the evaluator will need to consider designs for the evaluation, and sources, methods, and procedures for collecting the needed data or information. Let us address each of those decisions briefly.

Selecting Designs for the Evaluation

Designs specify the organization or structure for collecting data. The design selected generally has implications for the sources and methods of data collection. Therefore, the evaluator should consider what types of designs may be appropriate for each evaluation question and discuss related issues with the stakeholders. In this section, we will introduce the primary categories of design so they can be considered for the evaluation plan. In Chapter 15, we will describe each of these designs in greater detail.

Many evaluators conceptualize designs as descriptive or causal. When the evaluation question is causal, evaluators may consider using a design to establish causality. In the United States and in many western countries today, policymakers are quite focused on determining impact and many agencies emphasize randomized control trials (RCTs). (See "A Focus on Measuring Outcomes" in Chapter 2.) RCTs or experimental designs are certainly one method for establishing causality, but others exist as well. Quasi-experimental designs are a design approach to establishing causality when random assignment is not feasible or appropriate (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). Economists and policy analysts often use statistical techniques such as multiple regression and multivariate methods to control for extraneous variables in field situations (Tabachnick & Fidell, 2001). Case studies can also be used to make causal arguments (Yin, 2009). At this stage, evaluators should consider how important, and how appropriate, causal arguments are to the stage of the program and the decisions to be made and whether experimental design is the method that should be used to establish those causal links. If so, the appropriate design should be selected and the implications discussed with stakeholders. Evaluators should explore

whether conditions and/or data sources (for example, for interrupted time-series designs) will permit use of the preferred design.

More often the evaluation question is a descriptive one—to show a trend, to illustrate a process, to convey the status of something, or to describe and analyze a program, process, or procedure.[1] Time-series designs may be selected to show a trend, as reflected in the question, "Are high school graduation rates declining?" A cross-sectional design may be used to describe behaviors or opinions of a large group. For example, a cross-sectional design might answer the question, "How do parents in our district select their child's school?" for policymakers in a school district with many school choices, such as public, magnet, charter, or home schooling. A case study design may be used to describe the critical components of a successful child-abuse prevention program. Guba and Lincoln (1981) developed the term "thick description" to refer to the findings of intensive, descriptive case studies. Such thick descriptions can be most useful in informing stakeholders about what is actually happening in a program. Descriptive designs are commonly used in needs assessments and monitoring or process studies. They also can be useful in impact studies designed to determine whether participants' final performance is at the desired level or to describe performance at critical stages of the program. (See Spiegel, Bruning, and Giddings [1999] for an innovative evaluation of an assessment conference for teachers.) Many summative evaluations consist of a mix of causal and descriptive designs, to avoid a black-box solution that fails to describe connections between the clients and the program.

The evaluator and stakeholders should examine each question carefully to identify any important research design issues relevant to the question. Most evaluations use several designs or combinations of designs to address different questions. Nevertheless, it is important to consider design at this stage. Agreements may have to be reached on the availability of comparison groups, the appropriateness of random assignment, the time for collecting data from multiple sources, the selection of cases, the timing of measures, and other issues relevant to implementing the evaluation.

At this stage, the evaluator may be ready to specify the exact design to be used if the intent of the question is quite clear and the limitations and flexibility permitted in data collection have been explored. For example, if the question is simply to examine a trend—"Has the number of pregnant women receiving prenatal care in the first trimester increased over the last five years?" "Has the number of high school graduates pursuing education at community colleges increased over the last decade?"—it may be perfectly appropriate to designate a simple time-series design at this point. However, if there is some interest in exploring the whys of the trends, the evaluator might need to explore further to determine the extent to which case study or cross-sectional components should be added. In some cases, the evaluator may simply specify whether the design will

[1]The political push to examine outcomes can sometimes prompt stakeholders to fail to consider their real information needs. It is the obligation of the evaluator to help stakeholders consider the array of information that evaluators can provide and how they can use it.

be causal or descriptive, and then select a specific design when the details of information needs, feasibility of different designs and methods, costs, and timelines are further developed. Design concerns should, however, be addressed before the conclusion of the planning process.

Designating the appropriate design for each question enhances communication between the stakeholders and the evaluator and helps the stakeholders envision how the study will actually be implemented. Through learning the details of the design, the stakeholders can raise any concerns they have about data collection or issues that might constrain the study. Changes can then be made at that point rather than in the middle of data collection.

Identifying Appropriate Sources of Information

Each evaluation question requires the collection of information on at least one variable, if not more. For each variable, the evaluator and stakeholders consider who or what might be the source(s) for that information. For example, to answer the question, "Have the critical program activities been delivered as planned?" it was agreed that needed information would include descriptions of how counselors interacted with students in an evaluation of a high school advising program on post secondary educational choices (what to do after high school). The information sources for such items would typically include the counselors and students who participated in this interaction. Secondary sources (used as necessary to supplement or cross-check information and perceptions) might be the high school teachers and administrators and program records.

To answer the question, "What impact does the computer-based WANDAH program have on the writing performance of high school students in the Jefferson High WANDAH writing class?" let us assume that information was needed on one holistic measure (teachers' judgments of overall writing quality on one assignment) and one analytic measure (percent of *to be* verbs). The source of information for both would be the students in the WANDAH classes—and students in some non-WANDAH classes, if a comparison group were used. The source is not the teacher or the rater who counts the *to be* verbs; they only judge, score, or transmit information about writing performance, and that information obviously emanates from the students. The source is the group of individuals or the location of existing information that can answer the question.

Using Existing Data as an Information Source. Evaluators—and clients—sometimes overlook the fact that not every question must be answered by collecting original data. Evaluators should determine whether information relevant to any of the evaluation questions already exists in readily available form. For example, are there extant evaluation reports, status reports, or data collected for other purposes that might provide complete or partial answers to some evaluation questions? School records, for example, will contain information on students' attendance, disciplinary incidents and actions, grades, scores on state standards tests, demographics, and more. Before moving to collecting new information,

evaluators should always ask the client, program managers, and deliverers of the program whether there are existing sources that might meet the information needs. However, they also need to judge the appropriateness of these sources. Internal data, specific to the organization, may or may not be collected and organized in a valid and reliable manner. For example, data on test scores and grades are likely to be more accurate than information on student behavior.

Public documents and databases are another major source of existing information. Examples of such data include the reports developed by the U.S. Census Bureau (including the Census of Governments, the Decennial Census of Population and Housing, the monthly Current Population Survey, and Survey of Income and Program Participation); statistics collected by the U.S. Department of Labor and other federal departments; the *City-County Data Book*; and reports and databases of various state, local, and nonprofit organizations. Most state departments of education maintain extensive data on schools' performances on standards, enrollment, disciplinary incidents, and the like. Today, much data are available online.

Such data are typically intended to be used by others. Therefore, the information is generally collected in a careful, standardized fashion and is likely to be more reliable and valid than much internal existing data. Although it is reliable and valid for the purposes for which it is collected, such information may not be reliable and valid, or sufficiently sensitive, for the program evaluation at hand. The evaluator should learn about the manner in which the information is collected, the definitions of the constructs, the sampling methods used, and the timeframe and population sampled to determine whether the data will be appropriate for the current program evaluation. For many, if not most, programs, such data throw too wide a net to capture any changes a program may have caused. As Jane Davidson recently remarked, a city environmental program may have successfully reduced pollution in a nearby river, but water samples from the ocean would not capture that effect (Davidson, 2010)!

One word of caution: Just because data exist does not mean the data must be used. We have no sympathy for the evaluator who permits evaluation questions to be wrested into nearly unrecognizable form only so they can be answered by available information. Such distortion of an evaluation's intent is not excusable by claims of heightened efficiency. In such cases, the evaluator has committed what Michael Patton (1986) has called a "Type III error," answering the wrong question!

Commonly Used Information Sources. Within each evaluation study, information sources will be selected to answer the particular questions posed. Obviously, information sources may be as idiosyncratic as the related questions. As discussed previously, existing data are one important information source. But in almost all evaluations, some original data are collected. The most common sources are these:

- Program recipients (students, patients, clients, trainees)
- Program deliverers (social workers, therapists, trainers, teachers, physicians, nurse practitioners)
- Persons who have knowledge of the program recipients (parents, spouses, coworkers, supervisors)

Other frequent sources for original data include:

- Program administrators
- Persons or groups who might be affected by the program or who could affect its operation (the general public, future participants, organizations or members of interest groups involved in the program)
- Policymakers (boards, CEOs, elected officials and their staffs)
- Persons who planned or funded the program (state department officials, legislators, federal funding agency officials)
- Persons with special expertise in the program's content or methodology (other program specialists, college or university researchers)
- Program events or activities that can be observed directly

Policies That Restrict Information Sources. It is important to identify, early in planning an evaluation, any policies that may affect the collection of information. Many organizations, such as schools and hospitals, have Institutional Review Boards (IRBs) that must review and approve all research and evaluation projects taking place in their organization. Talk with people at the IRB and obtain copies of their procedures. They will also have copies of any organizational policies that will affect data collection. Some issues to consider include obtaining parental or guardian permission if original data are to be collected from children or adolescents. Children and adolescents must also agree to participate. If data are being collected from adults, of course, they too must provide consent. (See "Protections to Human Subjects and Institutional Review Boards" in Chapter 3.)

Organizations may have other policies that influence evaluators' choice of sources for information, methods, and procedures. For example, contracts or agency policies may restrict how employees can be involved in the evaluation. Employees may be restricted from data collection or other tasks beyond their immediate job responsibilities. Most organizations have policies concerning data collection from clients or existing files. Many organizations require that surveys or interview questions be approved prior to use. Constraints may also exist about the use of personnel information. Such policies are often designed to protect the interests of clients; however, the evaluator needs to be aware of such policies to learn how they may limit or restrict data collection.

Client Involvement in Identifying Information Sources. The client's role in identifying information sources is nearly as important as client involvement in determining what information is needed. The evaluator will often, by dint of experience, be able to identify good sources of information that might not have occurred to the client. Almost as often, the client will be able to identify useful sources of information that might otherwise escape the evaluator's attention. It is simple enough to ask the client, "Do you have any suggestions about where we might best obtain information on teachers' use of discussion groups?" This sort of collaboration not only yields helpful answers, but also further enhances the shared ownership of the evaluation by the client and evaluator.

Identifying Appropriate Methods for Collecting Information

Once the evaluator has specified where or from whom the needed evaluation information will be obtained, the next step is to specify the particular methods and instruments for collecting the information. Returning to our earlier examples, information about the delivery of critical program events might be obtained through a combination of observation, diaries or logs kept by program staff, and interviews or surveys of program participants. Information about the impact of the WANDAH program on students' writing ability might be collected by the previously mentioned holistic measure (teachers' judgments of overall writing quality on a given assignment) or analytic measure (percentage of *to be* verbs in one writing assignment).

There are countless ways to classify data collection methods and instruments. Although not exhaustive, we have found the following classification scheme to be useful in prompting neophyte evaluators' thinking about possible methods of data collection (Worthen, Borg, & White, 1993).

- **I.** Data collected directly from individuals identified as sources of information
  - **A.** Self-reports of attitudes, opinions, behavior, personal characteristics, or history
    - **1.** Surveys or questionnaires (structured or unstructured; administered on paper, by computer, or orally by telephone or in person)
    - **2.** Interviews
    - **3.** Focus groups
    - **4.** Personal records kept at evaluator's request (e.g., diaries, logs)
  - **B.** Personal products
    - **1.** Tests
      - **a.** Supplied answer (essay, completion, short response, problem solving)
      - **b.** Selected answer (multiple-choice, true–false, matching, ranking)
    - **2.** Performances (simulations, role-playing, debates, pilot competency testing)
    - **3.** Samples of work (portfolios, work products of employees)
- **II.** Data collected by an independent observer
  - **A.** Open-ended observations
  - **B.** Observation forms (observation schedules, rating scales, checklists)
- **III.** Data collected by a technological device
  - **A.** Audiotape
  - **B.** Videotape
  - **C.** Photographs
  - **D.** Other devices
    - **1.** Physical devices (body-mass index, blood pressure, air quality, blood-alcohol content, traffic frequency or speed)
    - **2.** Graphic recordings of performance skills
- **IV.** Data collected with unobtrusive measures

V. Data collected from existing organizational information or formal repositories or databases
  A. Records
    1. Information collected by agencies for use by others (Census information, Labor Department, State Departments of Education)
    2. Official records of an organization (attendance, personnel, proposals, annual reports, databases established for performance monitoring, client records)
  B. Documents
    1. Organizational documents (notes or products of employees or program deliverers, manuals, reports, audits, publications, minutes of meetings)
    2. Personal documents (correspondence or e-mail files, notes, lesson plans)

Reviewing the Adequacy of Methods for Collecting Information. Many evaluators choose data collection techniques or instruments more for their familiarity than for their appropriateness. Evaluators frequently find familiar techniques to be applicable, but, equally often, new approaches must be sought. The review of literature conducted during the planning phase can acquaint evaluators with different methods used for collecting information on a particular construct. Information collected for research may differ from the evaluation needs of a specific site. However, such literature can help evaluators consider new methods for collecting information or it can save them the time of developing a new measure when a good one, relevant to the needs of the stakeholders, already exists. As an example, one of us who was recently involved in a multi year study training middle school and high school counselors in new roles and responsibilities found a useful, validated survey on counselors' roles to be used with teachers and administrators during a literature review. Use of this survey will also allow us to compare the views in the schools in the evaluation with those of others who have used the survey.

In addition to making sure the information collected matches the construct of interest, the evaluator should ensure that sufficient information will be collected on each construct. Some phenomena are so clear-cut (e.g., height, number of children in a classroom, dollars spent) that only one measure is needed. Others, such as writing ability or parenting skills, require multiple measures because no one measure is sufficient to capture the totality of the phenomenon. In such cases, multiple measures, using different sources and/or different methods, are necessary to ensure that the evaluation question is answered completely.

Once information collection techniques have been specified for each evaluative question, the evaluator should review them, as a set, to assess their technical soundness, availability, relevance, and utility, by asking these questions:

- Will the information to be collected provide a comprehensive picture of the construct or phenomenon specified in the evaluation question?
- Will the information collected be reliable and valid for the purposes of the evaluation?

- Are the procedures for collecting the information legal and ethical?
- Will the costs of any data-collection procedure be worthwhile, given the amount and kind of information it will provide? (Consider costs to the evaluation, to the organization, and to the participants.)
- Can the information be collected without undue disruption?
- Can the procedures be carried out within the time constraints of the evaluation?

Of course, the evaluator responsible for a particular type of data collection should have competence in that method. Evaluators need a bigger bag of tools than researchers because they are examining a wider variety of phenomena than most researchers. Therefore, it is imperative for evaluators to frequently reassess the methods they know and to seek training concerning new techniques. If the evaluation requires a method or measure that is new to the team, the evaluation manager can consider using a consultant or hiring a new member if the method is a major one. Such considerations are a critical part of staffing the evaluation, as we will discuss later.

Role of the Client in Identifying Methods. Typically, the evaluator will have more expertise regarding the array of possible methods than will the client or members of the advisory group. However, it can be useful to involve the client or advisory group in the selection of methods to receive their feedback. These stakeholders can often provide a fresh perspective and insights on how the people who are having data collected from them might react. The wording of questions, the focus of the observation, the means for building rapport and alleviating anxiety in focus groups and interviews, or the feasibility of physical measures can all be elements for useful discussion. Finally, the methods for collecting information will form the foundation of the evaluation. If the client does not find the methods to be credible or believe they will yield useful information, more dialogue between the evaluator and the client is necessary. The evaluator may work to educate and persuade the client as to the validity of the methods or, if that fails, may choose other methods that will yield credible evidence for the client or other important stakeholders. (See, for example, Mark & Henry [2006] for a useful discussion of their views on why causal methods are more important to policymakers.)

In most instances, collecting the evaluation information is the province of the evaluator, not the client, for reasons discussed in earlier chapters (e.g., conflict of interest, technical competence). It is, after all, the evaluator who must guarantee the ultimate quality of the evaluation information—the core of the evaluation. While the evaluator has the responsibility to ensure that data-collection procedures are designed and implemented to ensure quality information, clients may be involved in the data collection. For transformative approaches or for capacity building, evaluators sometimes cede more responsibilities to stakeholders at this stage (Fetterman & Wandersman, 2005, 2007; King, 2007; Mertens, 2008). In any case, when others are involved, it is the evaluator's responsibility to ensure that they have the training to carry out the data collection in a responsible manner.

Determining Appropriate Conditions for Collecting Information: Sampling and Procedures

It is not enough to specify only the methods and instruments for collecting information. As noted previously, the evaluator must also ensure that the conditions within which those methods and instruments are employed are appropriate. Perhaps the most common concerns are these: (1) Will sampling be used in collecting the information? (2) How will the information actually be collected? and (3) When will the information be collected? A few words about each of these concerns may be helpful.

Specifying Sampling Procedures to Be Employed. In considering sampling procedures, evaluators need to decide, first, whether they need to draw a sample or need to collect data from the entire group of interest; if sampling, they need to determine the strategy they will use. Samples may be random, with a goal of identifying a sample that is representative of the population, or purposive, with a goal of selecting a particular group, such as those at risk of dropping out, those who have performed best, students who have recently moved to the school, or patients who quit coming to a clinic.

Some innocents have stated that researchers use random sampling because they are concerned with generalizing their findings to large populations, whereas evaluators do not use sampling procedures because they are concerned only with describing and judging what exists in a particular case. In fact, basic research rarely makes use of random samples because the population of interest is too large to make random sampling feasible. The primary occasion when we see random samples is in polls where the population, typically potential voters in the next election, can be clearly specified. Although evaluation is concerned with the particulars of the local context, random sampling is sometimes used by evaluators. In some large evaluations, sampling can both help evaluators generalize to the population of the program and make the evaluation more efficient in its use of resources.

For example, if an evaluation team were asked to evaluate the effect of the state health care plan in Massachusetts on health costs and patient health, it is unlikely that they would propose collecting information on every citizen of Massachusetts. The cost would likely be prohibitive and probably unjustified as well. Careful use of systematic sampling procedures would permit the team to select and collect data on a much smaller group while still generalizing with a high degree of confidence about the likely impact of the Massachusetts health insurance program. Similarly, no sane educational evaluator charged with evaluating the effect of a district-wide, middle school math curriculum on student achievement in a large metropolitan school district would propose new testing on every middle school student (though he or she might make use of data from existing tests if the test items reflect the goals of the curricula).

However, when the group of interest to the study is very small, it is advisable to try to collect data from the entire population. In such circumstances, sampling methods are not as likely to result in a sample representative of the population. For

example, in an evaluation of an employment-training program with 118 trainees in which the evaluator wanted to assess participants' learning, it would be advisable to administer a measure of learning to each of the 118 trainees enrolled in the program. This could easily be arranged if the measure were administered by evaluation staff during class sessions.

Purposive sampling is frequently used in evaluation just as it is in research. Purposive sampling does not mean drawing a sloppy sample, such as anyone who volunteers or is encountered in the data collection process. Instead, purposive sampling means drawing a sample *on purpose*. In this case, evaluators and stakeholders, have an interest in the opinions or performance of a particular subgroup. Common subgroups selected for purposive samples are poor performers, those not responding as desired to the program or policy, and best performers. The purposive sample of poor performers can help the evaluator learn more about their difficulties and provide information to improve the program for them or develop an alternative. Studying those who have or are likely to perform very well can also provide useful information. The idea with this strategy is that if the best cases are not able to achieve the program goals, it is unlikely that others will be able to so. (See Fitzpatrick and Greene [2001] for Greene's discussion of using this strategy in an evaluation of a training program to learn more about how those most likely to implement the conflict-resolution strategies taught in the program did so in the field.) When extensive data have to be collected on an issue, purposive sampling, with careful consideration and identification of those to be selected, can be more useful than a representative sample, particularly when the purpose is not generalization but exploration and learning. For example, intensive interviews are almost always conducted with a purposive sample, that is, those who know or have experienced much about the issue of concern.

Sampling, then, is a tool to be employed by the evaluator whenever resources or time are limited, when the group of concern is quite large, or when there is interest in a particular subgroup. Consider whether either of these sampling strategies are necessary to answer the evaluation questions and, if so, include them in the evaluation plan. These activities will have to be considered when developing the budget and establishing the timeline.

Specifying How Information Will Be Collected. For each type of data collection, it is necessary to specify who will collect the data as well as the conditions under which it will be collected. Issues to be considered include:

- Who will collect the information? Will they be evaluation staff, program deliverers, volunteers, or others? For methods such as interviews, observations, and focus groups, how will the characteristics of those collecting the information influence the behavior of participants?
- What training should be given to the people collecting the data? What sorts of checks should occur as data collection proceeds?
- In what setting should data collection take place? Is the setting conducive to the participants providing the desired information?

- How will evaluators obtain informed consent?
- In what procedures are anonymity or confidentiality necessary? How can these conditions be conveyed to the participants and carried out?
- Does the data collection pose any risks to the participants? What are the potential benefits to them? Will the data collection and its procedures "respect the stakeholders' dignity and self-worth" as recommended in the AEA Guiding Principles (see Appendix A)?
- Is any special equipment or material needed for the collection?

Specifying When the Information Will Be Collected. It seems almost a truism to say that evaluation information collected too late to bear on the relevant course of events is not useful. Timeliness is essential. In determining when information should be collected, evaluators should consider three criteria:

1. When will the information be needed?
2. When will the information be available?
3. When can the information conveniently be collected?

Knowing when information will be needed establishes the latest allowable date for collecting it, because time must be allowed to analyze, interpret, and report results. Availability is also an issue. It is patently absurd to schedule student post-testing for early June if the school year ends in late May, yet we have seen evaluators who have discovered this fact too late. Similarly, mailing surveys in mid-December when many people are too busy with holiday activities is not wise planning. It is also inefficient to return repeatedly to a site to collect data that could have been collected only once, given better planning. If the evaluator specifies the time for each data-collection technique, it is easy to see whether data pertaining to other evaluation questions might be conveniently collected at the same time.

Determining Appropriate Methods and Techniques for Organizing, Analyzing, and Interpreting Information

Evaluators must plan the format in which information will be collected in addition to designating the means for coding, organizing, storing, and retrieving it. Although computing and computer databases have greatly simplified this task, evaluators should still consider who will create and maintain the databases, how and with whom they will be shared, who will have access for analysis, and so forth. An example of failing to think ahead might underscore this point. A consultant of our acquaintance was once called by a school district to help analyze "some evaluation data we have collected and would like to analyze in the next week or two." After asking to see the data, our friend was led to a room nearly half the size of a normal classroom. There were the data—thousands of students' notebook diaries bound in bundles, by classroom and school, filling the room and stacked floor to ceiling, except for passageways. Our friend's first fear was that the

data might topple over on him; his second was that district officials might really believe that all that data could be analyzed adequately in such a short time. After some discussion with our friend, school officials realized that analyzing a random sample of the data was all that was possible. It also occurred to them that they could have greatly simplified the lives of the students and spared their time (not to mention the forests of the Northwest) if that had been all the data they had collected in the first place.

Specifying How Information Is to Be Analyzed. For each evaluation question, the evaluator should describe the way in which the collected information will be analyzed. This requires two steps: (1) identifying the statistical or summarizing techniques to be employed for analyzing the information, and (2) designating some means for conducting the analysis. For instance, in the example with the school diaries, prominent themes might be identified with qualitative software and quotes selected for the report to illustrate the nature of the diary entries on specific themes. Central tendency and dispersion descriptive statistics might be used to summarize the frequency and variability of the themes. The *means* might refer to identifying software that may be useful in analyzing the data or staff time to examine, code, or sort and enter diary entries. Software purchases and staff time must be considered for the management evaluation plan.

Interpreting the Results. Statistical tables or summaries do not speak for themselves. Different people looking at the same results may attach very different interpretations to them, depending on their values, past experiences, and personal expectations. For this reason, it is useful to share the results of data analyses with the evaluation client and other key stakeholders as they become available, to elicit their interpretations of what those results mean. For some evaluation questions, the criteria and standards developed will serve as a guide to the interpretations. However, the evaluation plan should allow time for different stakeholders to review the information, for dialogue between groups and with the evaluator, and for deliberation and inclusion of different perspectives as needed in written reports.

Determining Appropriate Ways to Report Evaluation Findings

For each evaluation question selected, the evaluator should specify when answers and interpretations should be prepared and for whom. For some questions, frequent periodic reports may be appropriate; for others, a single report may suffice. Some reports should be formal technical documents; others may take the form of memoranda, informal discussions, oral presentations, or meetings.

A good way to plan the reporting of evaluation findings is to use a matrix that specifies the following for each evaluation question: (1) the audience, (2) the content to be included, (3) the reporting format, (4) the date of the report, and (5) the context in which the report will be presented. An example is shown in Table 14.1.

TABLE 14.1 *Sample Work Sheet for Planning the Evaluation Reporting*

Evaluation Question	*Audience for the Report*	*Report Content*	*Report Format*	*Reporting Schedule*	*Context for Presenting Report*
1. Have the critical program events been delivered as planned?	Program managers	Progress to date; problems requiring attention	Memorandum and verbal presentation	Beginning of each month	Presentation at staff meeting, with one-page written summary
2. What impact does WANDAH have on students' writing ability?	School principal, language arts faculty, school board	Findings of student performance on holistic and analytic measures	Written report, with oral briefing, plus executive summary	Preliminary report on March 15; final report on May 1	Briefing and discussion of preliminary report with faculty and principal; written final report to them and executive summary to board, with oral briefing as requested by board

Once the evaluators have identified the reporting needs for each evaluation question, they should review the reports to see whether collectively they provide the needed information in a usable form. In Chapter 17, we discuss evaluation reporting at some length. At the planning stage, however, we cannot improve on a very useful set of questions suggested by Brinkerhoff and colleagues (1983):

1. Are report audiences defined? Are they sufficiently comprehensive?
2. Are report formats, content, and schedules appropriate for audience needs?
3. Will the evaluation report balanced information?
4. Will reports be timely and efficient?
5. Is the report plan responsive to rights for knowledge and information with respect to relevant audiences? (p. 48)

Work Sheets to Summarize an Evaluation Plan

It may be useful to summarize briefly our discussion of the items that collectively form the outline of an evaluation plan. For each evaluation question used to focus the study, it is important to specify the following:

1. *Information required* to answer the question (constructs or variables on which information will be collected)
2. *Design(s)* to be used to collect information
3. *Source(s)* of that information
4. *Method(s)* for collecting the information

TABLE 14.2 *Sample Work Sheet for Summarizing an Evaluation Plan*

Evaluation Question	*Information Required*	*Design*	*Information Source*	*Method for Collecting Information*
What types of employment did welfare mothers find after completing their employment training? Did it include health benefits? Was the compensation adequate to bring about self-sufficiency?	Job title, responsibilities, sector of employment (public, private, nonprofit), number of hours per week, salary, health benefits, length of time employed, other components as they arise	Descriptive, cross-sectional, possible case study elements	Graduates of employment training	Survey, interviews, possible focus group

Sampling	*Information Collection Procedures*	*Schedule*	*Analysis Procedures*
Survey to all (n = 50), interview with sample of 20; focus group with 10	Surveys distributed when client picks up graduation diploma; interviews arranged then and conducted in their home by trained research assistants; focus group conducted by consultant, 15 randomly recruited with compensation of $25 and babysitting provided	Surveys—October; interviews—November; focus group—early December	Descriptive stats and chi square for surveys. Use results for interview. Summarize major themes of interviews. Use results to plan focus groups. Use taped transcript of focus groups for analysis. Integrate all results to describe trends and solutions.

Reporting Procedures				
Interpretation Procedures	*Audience(s)*	*Content*	*Format*	*Schedule*
Are at least two-thirds of those who are employed earning a sufficient amount to sustain their family? Are they able to afford adequate child care? Are health benefits provided? For those whose employment does not establish or appear to lead to self-sufficiency, what solutions are recommended?	Funding sources (city and state departments), project administrators, program deliverers (especially employment counselors), clients, public at large	Help answer question: What is the program doing well? What changes are needed?	Technical report to funding sources and project administrators with one meeting with each funding source to discuss results; several meetings with project administrators and deliverers to discuss their interpretation of results and possible changes; meeting with clients to report results and receive their input; press release to general public	Meetings to discuss results—January; release report—mid-February with press release

5. *Information collection arrangements*, including
   a. Sampling procedure (if any)
   b. Collection procedure (who collects information; under what conditions)
   c. Schedule for collection
6. *Analysis procedures*
7. *Interpretation procedures* (including standards as appropriate)
8. *Reporting procedures*, including
   a. Audience(s) for report
   b. Report content
   c. Report format
   d. Schedule for reporting
   e. Context for reporting

An efficient way of completing these steps is to use a matrix with the first column listing the evaluation questions and subsequent column headings corresponding to each important element of the plan, as shown in Table 14.2. This table is comprehensive. Of course, it can be condensed using columns that are most important to a particular evaluation. We illustrate a somewhat simplified version in Figure 14.1. This version is especially useful with clients, who can more readily assist in completing this short form, and with funding agencies, which have found such matrices useful in understanding what is proposed by the evaluator. The evaluator can, of course, subsequently add columns and detail as desired, for managing the project. A simple device of this type is among the most useful tools an evaluator can employ for summarizing or communicating an evaluation plan to clients and other audiences.

Evaluation Questions or Objectives	Information Required	Information Source	Method for Collecting Information	Information Collection Arrangements			Analysis Procedures	Interpretation Procedures and Criteria	Reporting of Information		
				By Whom	Conditions	When			To Whom	How	When

This sample worksheet (reduced in size from the original) is suggested for use in compiling information needed to prepare an evaluation plan. When completed, all the essential ingredients of an evaluation plan or design are present and can be summarized to communicate key features of the plan to clients and coworkers.

FIGURE 14.1 ***Sample Work Sheet for Summarizing an Evaluation Plan: Abbreviated Form***

Specifying How the Evaluation Will Be Conducted: The Management Plan

A final task in planning the evaluation study is describing how it will be carried out. A management plan is essential to help in overseeing the project. Who will do what? How much will it cost? What are critical milestones? What is the timeline or schedule of events? Conducting a thorough and systematic evaluation study is a complex undertaking. To make the effort successful, the evaluator must effectively manage not only the evaluation activities but also the resources allocated to carry them out. Bell (2004) emphasizes the vital importance of good management to evaluation, writing, "A valid and useful evaluation depends as much on effective management as on elegant study design" (p. 603).

Evaluation management is multifaceted. An evaluation manager must do the following:

- Hire, train, motivate, and monitor the work of other project staff
- Serve as liaison to evaluation clients, participants, and other stakeholders
- Identify and cope with political influences
- Develop, negotiate, and monitor budgets and, if necessary, negotiate changes
- Develop and monitor the timeline and schedules for the evaluation to ensure that products are completed on time and negotiate schedule changes as needed
- Ensure that all activities meet the technical and ethical standards expected of a good evaluation

Managing an evaluation requires good personal, communication, and organizational skills. Good personal and communication skills are essential to supervising evaluation staff effectively and to communicating with stakeholders and other audiences. Organizational skills are required for developing timeframes and budgets, monitoring them effectively, and knowing when to negotiate changes. An evaluation, whether a team or single-person effort, cannot afford to be disorganized or haphazard. As the *Program Evaluation Standards* remind us, good evaluations involve using "effective project management strategies" (F1) and using "resources effectively and efficiently" (F4) (Joint Committee, 2010). A management plan is needed to structure and control resources, including time, money, and people. As is the case with the evaluation plan, the management plan must be open to change in response to fluctuating circumstances, but the need for flexibility in no way diminishes the need for a plan.

A good management plan must specify the following for each evaluation question: (1) the tasks to be performed and the timelines for each task, (2) the personnel and other resources required to complete the task, and (3) the cost. Each column builds on the next. Once tasks and timelines are specified, the personnel and resources to perform the tasks can be identified. Complex tasks will require higher-level personnel; time-consuming tasks will require more hours. When tasks are compressed into a short timeframe, more personnel will be required. Finally, with the specification of personnel and resources, costs for each question can be determined. Table 14.3 presents a sample management plan. The following sections will describe how it is developed.

TABLE 14.3 ***Sample Management Plan Work Sheet***

Evaluation Questions	*Tasks*	*Timeframe*
1. Have the critical activities of the program been delivered as planned?	1a. Work with agency staff and program developer to identify program's critical activities and key elements of their delivery; review literature on theories and actions of similar programs. Identify key elements to monitor.	1a. First month of evaluation
	1b. Identify methods for monitoring activities and plan data collection	1b. Second month of evaluation
	1c. Collect, analyze, and interpret data	1c. Months 2–4 (first months of program)
	1d. Meet with agency and stakeholders to discuss findings	1d. Ongoing, once a week, months 3 and 4
2. What are the most pressing needs of unemployed youth (18–25) in Lake City?	2a. Meetings with agency staff, review of existing documents, identify key informants and other sources of data	First month of evaluation
	2b. Interview key informants, conduct 4 focus groups, 3 with youth, 1 with those who interact or serve them	Months 2–3
	2c. Analyze and interpret results of interviews and existing data	Month 4
	2d. Meet with program administrators to review results	Last week of month 4

Personnel Involved and Estimated Costs	*Other Resources Needed and Costs*	*Total Task Cost*
1a. Lead evaluator, 2 days at $1,000 per day = $2,000	None	$2,000
b. Evaluation staff (1), 5 days at $500 per day = $2,500	None	$2,500
c. Evaluation staff, 2 days a month at $500 per day = $1,000; Lead evaluator, 0.5 days at $1,000 per day = $500 per month of program (3 months)	None	$3,000
d. Evaluation staff, 2 hrs a week for 2 months = 2 days at $500 per day = $1,000; Lead evaluator, 1 hr a week for 2 months for discussions with evaluation staff and periodic meetings with client = 1 day at $1,000 a day = $1,000	None	$2,000

(*continued*)

TABLE 14.3 ***Continued***

Personnel Involved and Estimated Costs	*Other Resources Needed and Costs*	*Total Task Cost*
2a. Evaluator, 1 day at $1,000 per day = $1,000; Evaluation staff, 2 days at $500 a day = $1,000	None	$2,000
b. Lead evaluator, 2 days at $1,000 per day = $2,000 (some interviews, recruit and direct focus group facilitators, observe 1–2 groups) Evaluation staff, 8 days at $500 per day = $4,000 (interview key informants, obtain, analyze, and synthesize existing data) Other staff, recruit for focus group, arrange for rooms, taping, refreshments, etc. 1 day @ $240 a day = $240	Focus group facilitator, 8 hrs for focus group and 8 hrs for prep and interpretation = 16 hrs or 2 days at $1,000 = $2,000; Reimbursement to focus group participants: 4 groups with 6 members, $30 each = $720; Rooms and refreshments for focus groups at $75 each = $300; Secretarial costs for transcribing tapes, 2 days at $160 per day = $320	$9,580
c. Evaluation staff, 3 days at $500 per day = $1,500		$1,500
d. Lead evaluator, 1 hour at $125 per hour = $125; Evaluation staff, 2 days at $500 per day = $1,000	Clerical time, 0.5 days at $240 = $120. Printing costs: $200	$1,445
e. Lead evaluator, 3 hours at $125 per hour = $375; Evaluation staff, 3 hours at $68 per hour = $204		$579

Estimating and Managing Time for Conducting Evaluation Tasks

Developing a timeline for conducting the evaluation tasks is critical to the evaluation plan. The first step in such a process is identifying critical end and interim dates. These may be dates when information must be received by important decision makers or dates when interim or final reports are due to funders. Then, one works backward from these dates. Among the important issues to be considered are these: What information is needed for such reports? How must the information be analyzed and interpreted and who must be involved? From whom will the data be collected? Are there organizational time constraints for the data collection? What measures have to be identified or developed to collect this information? Who should be involved in that planning? What literature reviews have to be conducted? What stakeholder groups

should be involved in the initial planning? As you can see, many questions are involved. But, the evaluator can specify key milestones (completion of the evaluation plan, finalizing data-collection measures and procedures, collecting various types of data, analyzing and interpreting it, developing report) and, then, identify the time periods needed for each to establish critical due dates.

Of course, initially making such estimates can be difficult. Consulting with others in the organization who have experience with similar data collection or the time required for planning tasks can be very helpful in making estimates. Evaluators new to managing an entire project often underestimate the amount of time needed to consult with others and the time needed for necessary dialogue and deliberation with stakeholders and among evaluation staff. But, such time is critical, both in the planning stages and as results are received and interpreted. Complex or lengthy tasks, or work completed by a new staff person, will require more milestones than other types of tasks, to monitor the activity and ensure timely completion. Milestones of monitoring a large project should not simply be the beginning and end points of data collection, but should include key interim steps during which the evaluation managers can discuss progress and findings with evaluation staff. Of course, some actions can occur simultaneously and the timeline can help the evaluation team leader identify those actions and consider how time can be best used to meet key milestones. Beginning and end times can then be identified for each task.

As more time is required for certain tasks, the evaluator should adjust the overall project timeline. This might be accomplished by reducing the time required for future tasks through adding more personnel, reducing the scope of work, or determining whether the timeframe for the study can be extended. The timeline is a tool, not a taskmaster. It gives the evaluation manager a means for organizing and monitoring progress, but as unanticipated events occur, such as unforeseen difficulties in data collection, the need for greater depth or additional collection of information due to ambiguities in results, changes in key stakeholders or program personnel, or a client's requests for additional information, the evaluator can and should adjust the timeline to meet these needs.

To lay out such projects, some evaluation managers make use of Gantt charts, which are simple displays that include proportionate, chronologically scaled timeframes for each evaluation task. A Gantt chart lists tasks on the vertical axis and a time scale on the horizontal axis. A horizontal line is drawn for each task to show how long it will take. An evaluator can look at a Gantt chart and tell at a glance when activities will begin and how long each will continue. Gantt charts are easy to prepare and, in addition to their management benefits, are effective in communicating evaluation plans. A sample Gantt chart is shown in Figure 14.2.

PERT charts are sometimes used to organize more complex projects and are often used by internal evaluation units to plan and coordinate related evaluations or the unit's evaluation work. PERT, which is an acronym for *Program Evaluation and Review Technique*, was developed by the U.S. Department of Defense as a management tool for multifaceted military projects. It has since been used in many other settings to examine the interrelationships among tasks and the time required

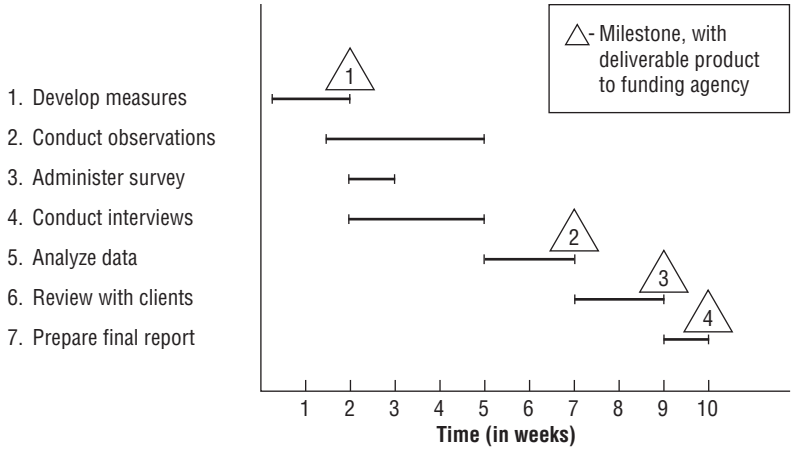
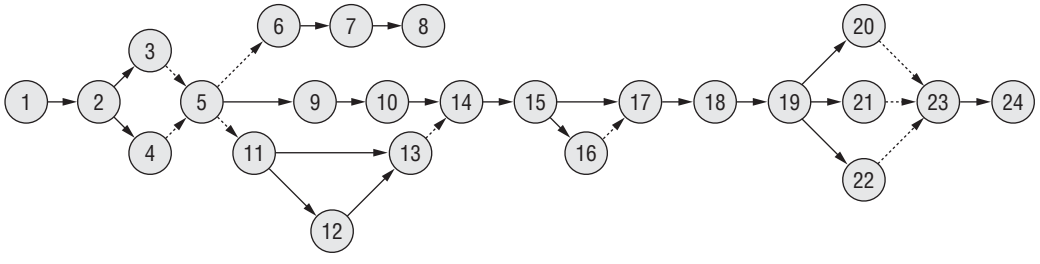


FIGURE 14.2 *Example of Showing Milestones on a Gantt Chart*



1. Start Project
2. Complete Objectives
3. Complete Data Paradigm
4. Complete Hypothesis
5. Start Item Construction
6. Start Universe Definition
7. Start Sampling
8. Start Sample Selection
9. Start Tryout
10. Start Final Form
11. Start Interviewer Selection
12. Complete Administration Procedures
13. Complete Schedules
14. Start Field Interview
15. Start Data Coding
16. Complete Follow-Up
17. Start Tabulation
18. Start Statistical Tests
19. Complete Tests
20. Complete Interpretation
21. Complete Tables
22. Complete Charts
23. Start Narrative
24. Complete Narrative

FIGURE 14.3 *Summary Network for Survey Research Project*

Source: From *Program Evaluation and Review Technique: Applications in Education* (Monograph No. 17, p. 43,) by D. L. Cook, 1966, Washington, DC: U.S. Office of Education Cooperative Research. Reprinted with permission.

to complete both subsets of tasks and entire projects (Cook, 1966; Sylvia, Meier, & Gunn, 1985). PERT charts are most useful in large, complex studies in which irresolvable problems might be created if details are overlooked. For many evaluations, however, PERT may be more cumbersome and time-consuming than enlightening. In most evaluation studies, a simplified version of PERT, in which one estimates the time required for a task and then links it with others that will be performed either simultaneously or before or after it, is sufficient. An example of a simplified PERT chart is shown in Figure 14.3.

Sufficient time must be allowed for all aspects of the evaluation—from focusing the study to final reporting. Build in extra time for the unexpected. For example, personnel turnover, among evaluation staff or with clients or key stakeholders, requires time for orienting these new people to the evaluation to ensure their good work, cooperation, and understanding. Good evaluation management should avoid placing key actors under unrealistic constraints that jeopardize the quality of their performance or stakeholders' participation and feedback.

Analyzing Personnel Needs and Assignments

The quality of any evaluation depends heavily on the capability and energy of those who carry it out. In some instances, only one individual—the evaluator—may be responsible for everything. Typically, though, others—members of the evaluation team, consultants, or clerical staff—will also be involved.

Perhaps the first concern of any evaluation manager is whether qualified individuals are available to carry out the various evaluation tasks. Addressing this concern requires specifying clearly the roles and responsibilities of each individual. In a relatively small evaluation, the lead evaluator must determine how she best spends her time. This will depend on her competencies, areas in need of leadership, and stakeholders' expectations, as well as the skills of other evaluation staff she might use. Bell (2004) suggests developing a staffing matrix with a list of tasks and the best person or skills required to perform that task. Obviously, the list would include tasks for the lead evaluator and other staff who are either employed by the organization or used frequently as research assistants on an interim basis. Such a list could make use of the columns shown in Table 14.3, or it could be an expanded list of the tasks and the individual(s) who will perform each.

To determine who will do what, consider the skills required for completing each task. Who has these skills? Who has experience and interest in this area? Who has training and experience in content analysis? In recruiting participants for focus groups? In analyzing data using path analysis? In writing clear, interesting reports for the public? For a small project, choices may be among the evaluator and one or two assistants or consultants. For larger projects, the skills of the existing professional staff, and their existing workloads, should be reviewed to determine who is available and most appropriate for each task. Even with a large existing staff, consultants may have to be hired in specialized areas.

For internal evaluators, many choices exist among the current organizational personnel who perform different duties. Consider how publications editors or public relations personnel can be of assistance in disseminating evaluation information. Computer and technology personnel may be able to provide assistance in online data collection, accessing and storing data, purchasing new software as necessary, and data analysis. Administrative assistants and finance staff may be used for routine tasks. In many participatory—and certainly in empowerment evaluations—program personnel and clients are involved in collecting, analyzing, and interpreting data. Consider the relevant skills, strengths, and risks each group or individual brings to the evaluation and use them accordingly. But, in some, if not many cases, others will have to be recruited to perform certain tasks.

The lead evaluator or the evaluation manager is responsible for recruiting, screening, and hiring evaluation staff for these tasks. Consider the evaluator competencies listed in Chapter 11 in Figure 11.3 and consider those that are pertinent to perform the tasks required for the position. Use local or state evaluation associations to recruit potential staff or consultants. (In the United States, see http://www.eval.org/aboutus/organization/affiliates.asp for a list of American Evaluation Association local, state, or regional affiliates who may post an announcement to their members or use "Find an Evaluator" at http://www.eval.org/find_an_evaluator/consultant_listing.asp. Professional evaluation associations in other countries or regions are likely to have similar means for recruiting or recommending evaluation professionals in the region.) Of course, once screened and hired, the lead evaluator is responsible for orienting and training the newly hired staff member concerning the organization and the project to be evaluated.

Estimating Costs of Evaluation Activities and Developing Evaluation Budgets

There are many nonmonetary and indirect costs of evaluation, such as opportunity costs or political costs. In the interest of simplicity, we limit our discussion here to direct dollar costs.

Typical Resources and Their Costs. An evaluation budget usually includes the following ten categories:

1. *Evaluation staff salary and benefits.* Personnel costs typically consume the largest part of the evaluation budget. The amount of time that staff members must spend on evaluation tasks and the level of expertise needed to perform them both affect costs. Decisions must be made about who will perform various tasks and the amount of time the tasks will require. Benefits have become an increasingly costly portion of the budget as health insurance costs have risen. Costs in this category are estimated relatively easily by using existing salary and benefit figures. Once the proportion of a person's time devoted to the evaluation is determined, that portion of the staff member's salary and benefits can be charged to the evaluation budget. Most organizations have a set benefit rate that is a percentage of the salary. If a new person must be hired, salaries for the proposed position can be determined by consulting with other organizations with like employees, perusing advertisements and notices in publications advertising for similar positions, and so on.

2. *Consultants.* As noted, consultants are sometimes needed either (a) to provide skills not currently reflected among project staff and not permanently needed in the organization, or (b) to provide an independent perspective on the program or the evaluation. Consultants also have the advantage (to the budget, at least!) of not receiving benefits. Costs for consultants can be calculated using their daily or hourly rate.

3. *Travel and per diem (for staff and consultants).* Costs depend on the amount of field work and the degree of personal interactions required to design and conduct

an evaluation. Some contracts apply restrictions on travel costs (e.g., no billing for travel within the catchment area of the organization). Travel costs can include estimates of automobile mileage for meetings, training, observations, data collection, and other activities outside of the catchment area. Airfare, ground transportation, and per diem costs for lodging and meals should be calculated for long-distance travel.

4. *Communications (postage, telephone calls, etc.)*. This category includes both fixed costs (e.g., continuing monthly billings for telephone lines, computer networking, and Internet access) and variable costs (for special communication efforts, such as conference calls, faxes). Fixed costs can be budgeted by multiplying the length of the contract by the proportion of work the organization devotes to that contract. (Many of these elements may already be counted in indirect or overhead costs.) Variable costs should be estimated based on the nature of the tasks involved. Therefore, postage costs will be much higher if mailed surveys are part of the study. These postage costs should be calculated directly based on the number of mailings, postcards, and return envelopes.

5. *Printing and duplication.* Costs cover preparation of surveys, protocols for observations and interviews, evaluation reports, and any other documents. Routine costs may be estimated by comparison with past projects or discussions with others who have overseen similar projects. Often clerical staff can be helpful in estimating budget costs in this area. Costs of printing, duplicating, and binding final reports, or any special graphics should be checked with copying centers. Print production costs can be reduced or eliminated if disseminating reports via e-mail or posting them on the Internet will reach the intended audience and attract their interest as well as other methods. However, costs for associated web-design issues should be considered and included as necessary.

6. *Computer costs.* Consider whether new software will have to be purchased for data analysis, storage, or retrieval. If web-based surveys are to be used, consider costs for contracting with an existing survey system, such as Survey Monkey, or whether staff with the necessary skills are available and a better choice to develop a web-based survey for the organization. Determine how the organization's web site, the evaluation company's site, the organization that manages or funds the program's site, or all three will be used to post evaluation reports and whether special web site management issues are necessary either for those reports or for communicating among evaluation staff and stakeholders.

7. *Printed materials.* This category includes the costs for purchasing existing data-collection instruments and library materials. Publishers of books or measures can provide costs for possible budget items.

8. *Supplies and equipment.* This category covers the costs of specific supplies as well as equipment that must be purchased or rented. If the primary supplies to be used are routine (pencils, pens, paper, etc.), typical office estimates should be obtained and prorated for the length of the contract. Occasionally, special purchases or rentals

are necessary. These could include videotaping equipment, specialized software or hardware required for the project, purchases of existing databases or fees for using existing data, or mechanical devices for collecting data (such as blood-pressure monitors). Costs of purchasing or renting such special equipment should be obtained from suppliers. In some cases, contracts may require rentals rather than purchases of costly equipment.

9. *Subcontracts.* This category includes expenditures for any contracted services, such as accounting, legal services, test development, and so forth. All subcontracts must be negotiated with the subcontractor before the evaluation budget is completed. Each subcontractor may submit an independent budget. Agencies and institutions often include these costs in their overhead rates. However, small or new agencies may need to bill for these services.

10. *Overhead (facilities, utilities).* The greater the use of external personnel and services, the lower the overhead costs. Typically, however, an institution must bear certain fixed overhead costs (i.e., those of maintaining an adequately equipped physical plant) regardless of what arrangements are made for the evaluation. Most organizations have fixed percentages of a total budget, or of personnel salaries and benefits, that they charge as operating overhead. Check to see what overhead covers to make sure you are not double-billing. If overhead includes fixed costs for communication, computers, accounting, or legal services, these should not be billed separately.

Once costs in each budget category have been calculated, a total cost for the evaluation can be determined. This first estimate often exceeds the evaluator's or client's expectations. When this happens, review each line item and ask how the work could be completed at less cost. Some effective cost-saving measures include:

- Using available volunteers or low-cost workers to reduce staff salaries and benefits
- Using local specialists for data collection to reduce travel costs if evaluation staff are at a distance
- Training less-costly personnel to perform selected tasks
- Borrowing (equipment, people, materials, and supplies)
- Seeking in-kind contributions from the organization in which the external evaluator is employed (often done for good public relations) or the sponsoring agency
- Reducing the scope of the evaluation, perhaps deferring some parts for the future
- Using existing measures, data, or reports
- Using inexpensive data collection when precision can be sacrificed without severe consequences
- Using public media to disseminate results
- Piggybacking on other studies
- Increasing efficiency through good management

Establishing Evaluation Agreements and Contracts

Many potential problems that arise during evaluation can be more readily solved if client and evaluator have a clear understanding of their own and each others' roles and responsibilities in regard to the evaluation. Even among administrators and evaluators with the highest possible professional standards and ethics, conflicts can and do arise because of unstated assumptions of expectations. Evaluators must discuss these issues with a client during the planning stage and develop an agreement that can be used to document expectations and responsibilities. Guba and Lincoln's (1981) cautions about evaluation contracts are still relevant today:

> Evaluations are done for clients who commission the evaluation, provide for its legitimization, and pay for it. Since he who pays the piper calls the tune, the evaluator must have a firm understanding with the client about what the evaluation is to accomplish, for whom, and by what methods. The evaluator also needs to be protected against certain arbitrary and possibly harmful or unethical actions by the client, just as the client needs to be protected against an unscrupulous evaluator. The means for achieving these understandings and establishing these safeguards is the evaluation contract. (pp. 270–271)

The Joint Committee's *Program Evaluation Standards* includes Formal Agreements as a specific standard: "P2: Evaluation contracts should be negotiated to make obligations explicit and take into account the needs, expectations, and cultural contexts of clients and other stakeholders" (Joint Committee, 2010). They suggest guidelines for the agreement and note common errors in drafting such agreements. Stufflebeam (1999) has developed a checklist for evaluation contracts that serves as a useful guide (see http://www.wmich.edu/evalctr/checklists/contracts.pdf). The checklist reminds us to include agreements on the information to be obtained and how it will be analyzed, reporting requirements, and expectations for both the evaluator and the client. Some of the thorny issues that should be resolved up front concern the ownership of the evaluation and its reports: Who has editorial authority? What are the review rights of the client and other stakeholders? Who has the final authority to release reports? Does the evaluator have permission to publish results of the evaluation or other issues concerning it in professional journals? These agreements, of course, should be negotiated during the planning stage, and a written agreement should be shared between the evaluator and the client or sponsor (or both if they differ). Finally, if the evaluation takes place over a relatively lengthy period of time, for example more than a year, the parties to the agreement should review it each year and renegotiate the agreement as needed.

Another agreement between clients and evaluators that should occur at this point concerns ethics and standards. The evaluator should share the Guiding Principles, the *Program Evaluation Standards,* or both with the client and other critical

stakeholders such as an advisory committee.[2] This information can help the audiences for the study know what to expect from the evaluation. Both the Guiding Principles and the *Standards* can be obtained at www.eval.org, or brochures may be ordered from that site for distribution to clients. (See Chapter 3 for more on the Guiding Principles and the *Program Evaluation Standards,* and Appendix A for a list of the key sections of both.)

Planning and Conducting the Metaevaluation

The agreement or contract specifies the stages of the evaluation itself, but the final stage of the planning process is considering when and how the evaluator, or the client, will seek input on the quality of the evaluation. Evaluators have come to recognize that they must evaluate themselves. We cannot simply evaluate others and tell them how to improve or what to continue, expand, or terminate without evaluating our own work. Metaevaluations were created to address this need. Evaluators realize that they may unwittingly introduce biases or errors into their work. They may ignore certain audiences or overlook cultural differences that hinder accurate data collection and communication. Both evaluators and clients must be concerned about the quality of the evaluation work: evaluators because their personal standards and professional reputation are at stake; clients because they don't want to invest (either politically or financially) in findings that are off-target. Both have a lot to lose if an evaluation is shown to be deficient in some critical aspect. This is why metaevaluation—the evaluation of an evaluation—is important. Formative metaevaluation can improve an evaluation study before it is irretrievably too late. Summative metaevaluation can add credibility to final results.

The Development of Metaevaluation and Its Use Today

In an informal sense, metaevaluation has been around as long as evaluation, for someone has had an opinion about the quality of every evaluation study ever conducted. During the 1960s, however, evaluators began to discuss formal metaevaluation procedures and criteria, writers began to suggest what constituted good and bad evaluations (e.g., Scriven, 1967; Stake, 1970; Stufflebeam, 1968), and unpublished checklists of evaluation standards began to be exchanged informally among evaluators. In addition, several evaluators published their proposed guidelines, or metaevaluation criteria, for use in judging evaluation plans or reports (Scriven, 1974b; Stake, 1969; Stufflebeam, 1974; Stufflebeam et al., 1971). At this early stage in the field of evaluation, there seemed to be little consensus about

[2]Readers in other countries should consult ethical codes for evaluation in their country. Canada, Europe, individual countries in Europe, and other countries around the world have developed ethical codes for evaluation.

which set of proposed metaevaluation criteria was preferable. There was, however, a desire to develop such a touchstone, because it seemed likely that a widely accepted set of criteria for determining the quality of an evaluation would lead to more and better metaevaluations.

In 1975, Daniel Stufflebeam chaired a Joint Committee on Standards for Educational Evaluation composed of 16 people appointed by their professional associations. Their task was to come to a consensus on such criteria. This Joint Committee produced the first set of evaluation standards designed specifically for the evaluations of educational programs, projects, and materials. Their original work comprised the first standards to be used by the professional community to guide and evaluate educational evaluations (Joint Committee, 1981). The Joint Committee has continued its work over the years and has been composed of representatives from 12 to 15 professional associations concerned with evaluation, including the American Evaluation Association, the Canadian Evaluation Society, the American Educational Research Association, the American Psychological Association, and many professional associations concerned with school administration and operations (Stufflebeam & Shinkfield, 2007; http://www.jcsee.org/). Their work has been quite successful. The *Standards* have expanded from their original focus on educational evaluation to apply to all program evaluations and have been accredited by the American National Standards Institute. They have been used extensively in the United States and were recently revised for a third edition (Joint Committee, 2010). Metaevaluations today often make use of these internationally accepted standards or the AEA Guiding Principles to judge evaluations.

In the second edition, in 1994, the Joint Committee added metaevaluation as a standard. The third revision of the *Standards* (2010) emphasized the increasing importance of metaevaluation by creating a new category, "Evaluation Accountability," with two standards concerning metaevaluation:

> **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. (Joint Committee, 2010).

Internal metaevaluations are conducted by the evaluators themselves using the *Standards*, the Guiding Principles, or other applicable standards to judge their work and, typically, to make changes to improve the quality of the evaluation as it is being conducted. An external metaevaluation, like an external evaluation, is conducted by an external party and may be used for formative or summative purposes. The external party may be a sponsor, a client, or another evaluator. The success of this approach depends heavily on the technical competence of the consumer to judge how well the evaluation meets such standards as "Valid Information" or "Sound Designs and Analyses." The Joint Committee *Standards* and the

AEA Guiding Principles, however, do not necessarily require specialized technical training. It may be quite feasible for a client to apply many of these criteria effectively, calling on an independent, external evaluator to apply those standards or principles that are more directly concerned with technical issues.

Unfortunately, metaevaluations are not conducted as frequently as they should be or, at least, few examples have appeared in the evaluation literature. But in recent years, a few useful examples have emerged. Hanssen, Lawrenz, and Dunet (2008) describe a metaevaluation they conducted, which they call a "concurrent metaevaluation" because their activities as metaevaluators were concurrent with the evaluation being reviewed. They were hired to provide feedback to the evaluators to improve a large evaluation as it was taking place. (We would consider it an example of a formative metaevaluation. But, as they note, most metaevaluations take place after the evaluation has been completed and, therefore, are not intended for improvement, but for judging the quality and the strengths and weaknesses of the evaluation. Their emphasis is conducting the metaevaluation concurrently with the evaluation to permit it to serve formative purposes.) The metaevaluation conducted by Hanssen and his colleagues is unusual in that they, the metaevaluators, were intensely involved with the evaluation to be reviewed. As in an evaluation study, they developed questions for the metaevaluation and collected data for each. For example, they participated in weekly evaluator team meetings, reviewed draft documents, observed site visits and interviewed a number of people who either collected data or provided data for the evaluation. Their feedback was intensive and productive in improving the evaluation. As with many metaevaluations, they made use of the *Standards* to conduct the metaevaluation, but added two other criteria, Scriven's *Key Evaluation Checklist* (2007) and the Society for Prevention Research's *Standards of Evidence* (2004). (The evaluation concerned health prevention activities at work sites.) Thus, their project illustrates the use of both additional standards as appropriate to a particular project and an intensive, formative metaevaluation that positively affected the evaluation they reviewed.

Another metaevaluation illustrates a quite different approach. Perry (2008) describes what she calls a "mental metaevaluation" she undertook using the *Standards* and AEA's Guiding Principles to review and evaluate an evaluation conducted by students in her course. Her article illustrates what can be learned by using some standards to consider an evaluation. She describes the evaluation and its context and, in assessing it, indicates particular *Standards* and Guiding Principles that helped her analyze particular issues in the evaluation. The metaevaluation helped her not only to analyze the particular evaluation, but also to consider her practice and that of her students for the future. She recommends having students conduct metaevaluations of projects they conduct in class to become accustomed to using the process to assess their own work. These two articles represent two quite distinct applications of metaevaluations, differing in cost, purpose, and the stage of the evaluation at which it was conducted, but each illustrates how metaevaluations can serve a quite useful purpose.

Patton (2008a) and Stufflebeam have both indicated that full metaevaluations are not necessary for every evaluation. If they are costly, they should be reserved for

evaluations that are what Patton calls "high stakes." He proposes the following guideline:

> The higher the stakes for evaluation use (e.g., summative evaluations), the more politicized the context in which an evaluation is conducted, and the more visible an evaluation will be in that politicized environment, the more important to credibility will be an independent assessment of evaluation quality. (2008a, p. 537)

Note that this guideline is certainly appropriate when a metaevaluation will require extensive resources, whether in actual dollars or in personnel use and time. But Perry's metaevaluation illustrates well how the metaevaluation can be used in a low-cost manner by either the evaluator or a colleague to assess and learn from an evaluation and thereby to improve future practice.

Some General Guidelines for Conducting Metaevaluations

Evaluators would do well to build on the work illustrated by Perry and Hanssen and his colleagues mentioned previously. Consider whether the metaevaluation is intended to improve the evaluation, in which case it should be conducted concurrently with the evaluation, or whether its purpose is to judge the quality of the evaluation for other audiences. Metaevaluations can be conducted at many different times for different purposes: after an evaluation plan or design has been finalized, at periodic intervals during the evaluation to check progress and identify problems, and at the end of the evaluation to review findings and reports and to audit the evaluation procedures and conclusions. Having defined the purpose and the timing of the metaevaluation, evaluators—or clients or sponsors if they have requested the metaevaluation—can then determine whether to use internal and external reviewers.

The internal review can be conducted by an evaluation colleague, or by an evaluation committee or advisory group. While the evaluation is in progress, the evaluator could enlist a group of stakeholders and evaluation staff and ask for their reactions to the evaluation plan, its implementation, the relative timeliness and costs of various tasks, and the need for any revisions. The minutes of such meetings provide useful progress reports for the client.

The external review is best conducted by a disinterested outside party with successful experience in similar evaluations. As demonstrated in the Hanssen et al. (2008) example, if called in early enough, the outside evaluator can review the evaluation plans and the actions being taken to implement it and offer recommendations for strengthening it. An external reviewer can also provide technical assistance during the evaluation and, at the end of the project, can review the results, conclusions, and reports. The external reviewer may need to schedule a site visit at each review stage to gain full access to evaluation files, instruments, data, reports, and audiences. Such an arrangement takes both planning and knowledge of how and where to access pertinent evaluation information and, of course, knowledge of different standards, their meanings, and their application.

Whoever is conducting the metaevaluation must then identify the standards to be used and the issues of focus—if not the entire evaluation. Typically, the metaevaluators would work either with the evaluator or evaluation team or with the client requesting the metaevaluation to learn its purpose and focus and would include them in discussing the standards to be used. The metaevaluator would then identify the information needed to conduct the metaevaluation. This often consists of a combination of existing documents and records (evaluation plans, minutes of meetings, databases, and evaluation reports) and people to interview, including the evaluators, clients, and stakeholders. Such information would be reviewed or collected in interviews and then analyzed, comparing it to the selected standards. Results would be synthesized and the metaevaluator would draw conclusions and recommendations as appropriate for the stage of the evaluation.

An abbreviated example may be of use here. We will focus on a metaevaluation of the choice of an evaluation design and its particulars. In evaluation, design is critical. Poor designs do not lead to satisfactory evaluations. We will assume that the metaevaluation was contracted for formative purposes, to provide feedback to the evaluator for design improvement and that the metaevaluation begins as the evaluation plan is completed. The metaevaluation would examine the design as planned and implemented. A complete metaevaluation for the design would include the following:

- Reviewing the proposed design to ensure that it is feasible and sound
- Monitoring the design to see that tasks are completed as planned and within budget
- Checking the quality of instruments, procedures, and products (such as data and reports)
- Reviewing the design for possible midstream revisions (especially in light of either the utility the evaluation has shown so far for important audiences or the problems the evaluation was encountering)
- Checking the effects of metaevaluation on the design of the evaluation

Because of limited space, we will restrict this discussion and our remaining examples to evaluating the design. Readers should easily be able to extrapolate the steps from this discussion to metaevaluations of other aspects of an evaluation.

Steps to Take in Evaluating an Evaluation Design. The following steps are proposed for conducting a metaevaluation of an evaluation design:

1. *Obtain a copy of the design in a form ready for review.* A formative metaevaluation of the design can be useful once the design is sufficiently formulated to make such a review productive, but before data collection begins. At this stage, the design can be modified based on feedback from the metaevaluation.

2. *Identify who will do the metaevaluation.* An internal evaluation unit might use an evaluator not assigned to this project. For a small, external evaluation, an evaluator might recruit an evaluation colleague who will have sufficient independence to

provide constructive feedback for changes. In a large external evaluation, an advisory council with the assistance of a technical consultant might conduct the metaevaluation and provide feedback. This would permit the evaluation team to receive input from clients about the utility of the design, feedback on its technical adequacy from the consultant, and response from the advisory group.

3. *Ensure that authorization exists to evaluate the design.* If you are a sponsor or client and you receive a design submitted by an evaluator who proposes to contract with you to do an evaluation, you are free to evaluate it, and normally there would be no professional or legal restraint on your arranging for another competent metaevaluator to assist you in doing so. Conversely, suppose the chair of a Concerned Citizens against Homeless Shelters committee asks you to find flaws in an internal evaluation design that the local homeless shelter proposes to use in evaluating its program. You should question the appropriateness of that role, especially if you find that the design is in rough-draft form, circulated only for internal reactions, and surreptitiously spirited from the shelter to the committee by a disgruntled employee. Metaevaluators (like evaluators) can find themselves being used as hired guns, and it is important before buckling on the holster to be certain that the metaevaluation desired by your paymaster will not violate ethical or legal principles.

4. *Select the standards to be used and apply them to the evaluation design at various stages of the evaluation.* The *Program Evaluation Standards* are often used to conduct a metaevaluation, but, as the *Standards* themselves express, other appropriate standards can be used. To conduct a metaevaluation of just the evaluation design, the metaevaluator might choose to use a combination of the *Standards* and the Guiding Principles that are relevant to design. For standards, the primary focus would be on A6, which explicitly concerns design, but other relevant standards include U3, P1, P3–P5, and A1–A5. Given the design selected and the involvement of stakeholders, the metaevaluator should also consider how the design and its selection conforms to standards U2, U4, U6, F1–F4, and E1. Guiding Principles that are most pertinent to a metaevaluation of the design are A1, B1, C6, D2–4, and E2. Other Guiding Principles that have relevance include A2, B2–3, C1, C4, D5–6, and E1, 4–5. (See the *Standards* and Guiding Principles in Appendix A.)

5. *Judge the adequacy of the evaluation design.* No evaluation design is perfect. The question is whether, on balance, after summarizing judgments across scales, the evaluation seems to achieve its purposes at an acceptable level of quality.

A Need for More Metaevaluation

With any luck, we have convinced the reader that the concept of metaevaluation is useful and that there are appropriate tools that can be used for that purpose. Despite the wide publicity, acceptance, and availability of the Joint Committee's *Standards*, however, few metaevaluations are being conducted. Of the few metaevaluations that do occur, most are internal evaluations done by the evaluator who produced the evaluation in the first place. It is rare indeed to see evaluators

call in an outside expert to evaluate their efforts. The reasons why this is so are many and complex, but one seems particularly compelling: Evaluators are human and are no more ecstatic about having their work evaluated than are professionals in other areas of endeavor. It can be a profoundly unnerving experience to swallow one's own prescriptions. Although the infrequency of good metaevaluation might be understandable, it is not easily forgivable, for it enables faulty evaluation practices to go undetected and, worse, to be repeated again and again, to the detriment of the profession.

Major Concepts and Theories

1. Evaluation consists of the following major functions: focusing the evaluation; collecting, organizing, analyzing, and reporting information; and administering the evaluation.

2. During the planning phase, the evaluator should begin to consider and develop the evaluation plan that will be used to document the evaluation questions to be answered and the methods for doing so. The evaluation plan should list the evaluation questions to be addressed by the study and, for each, the design(s), data sources, methods, procedures, analyses, and methods of interpretation to be used for each question. This plan is a key document for communicating the purposes, focus, and means of conducting the evaluation with the client and key stakeholders.

3. A second plan is needed to help the lead evaluator manage the evaluation. A management plan should be developed specifying the tasks to be completed, and the timeline, personnel, and costs associated with each task. The management plan serves as a guide for monitoring and overseeing the evaluation but should be adapted for changing circumstances. Managing the evaluation also includes assigning and possibly hiring and orienting staff for the project, as well as developing and monitoring the budget for the evaluation.

4. In developing the budget, the evaluator needs to consider evaluation staff salary and benefits, consultants, travel and per diem, communication, printing and duplication, supplies and equipment, and overhead. Each of these is dependent on the type of evaluation and data-collection methods used, as well as the extent to which stakeholders will be involved in each task and the nature of the evaluation reporting that will be carried out.

5. The evaluator should work out a contractual agreement with the client that indicates clearly the purpose of the evaluation, the activities to be completed, and the responsibilities of the evaluator and the client.

6. Finally, evaluators should make use of metaevaluations, as necessary, to receive feedback from others to improve the evaluation or to judge the quality of the evaluation. Metaevaluations typically make use of the *Program Evaluation Standards*, AEA's Guiding Principles, or other appropriate standards as criteria for conducting the metaevaluation.

Discussion Questions

1. Which type of sampling do you think is most feasible and most useful in most evaluations—random sampling or purposive sampling? When might an evaluator productively use each strategy?
2. What are some ways evaluators can decrease the costs associated with an evaluation?
3. Not surprisingly, some methodological choices prompt much discussion and debate among evaluators. For example, Mark and Henry (2006) argue that policymakers are most interested in and most likely to use evaluation results that establish the causal connections between a program or policy and an outcome. Others argue that policymakers don't really use such results, that they're too busy with political pressures and have their own opinions about what works. These evaluators argue that evaluation is most likely to be used at the program level where program managers and staff are often more interested in descriptive information. Which type of designs (causal or descriptive) do you think are more likely to be used? By whom? Why?
4. Why is a contractual agreement between the evaluator and the client useful? In what areas do you think disputes are most likely to arise if a contract is not developed and reviewed? (See Stufflebeam's evaluation contract checklist at http://www.wmich.edu/evalctr/checklists/contracts.pdf to facilitate this discussion.)
5. Read Perry's article on her metaevaluation of an evaluation conducted by her students and discuss whether you would draw the same conclusions. Discuss, also, the feasibility of using the process she describes with evaluations you know. Would such a process be helpful? In what ways?

Application Exercises

1. Using the evaluation questions you developed at the conclusion of Chapter 13, develop an evaluation and management plan to address those questions. What further information do you need to do that? (Subsequent chapters will tell you more about methods.) What stakeholders should you involve in planning the evaluation design? What tasks are involved in each step? Who should do them? When?
2. Select one of the case study interviews we have recommended at the end of each chapter. Reconstruct an evaluation plan from that article. What were the evaluation questions the study answered? What information was collected? What designs, sources, and methods were used? Were multiple methods used? How were data analyzed and interpreted? How were stakeholders or clients involved in planning the evaluation?
3. Interview someone who has conducted an evaluation study. Ask her how she developed her design. What issues were most troublesome at the planning stage? How did she involve stakeholders on the issues? On which issues did stakeholders play a major role? On which did the evaluator hold more decision-making power? Why? How would you have developed the study differently? What aspects of her plan would you now incorporate into your own matrix?

4. During your interview for exercise 3, ask how the evaluator managed the study. What did professional evaluators do? Research assistants? Clerical staff? What kinds of tasks were accomplished by internal staff? What was their connection to the program and how were they trained? How did the lead evaluator monitor the timeline? Finally, see if she will give you a copy of the budget and ask her how she determined some of the costs. Did costs change as the project developed? Bring the budget to class and compare it with budgets obtained by other students.
5. If your class is conducting an evaluation, try Perry's strategy and use the *Standards* and Guiding Principles to conduct a mental metaevaluation on your project. Although Perry's metaevaluation occurs after the evaluation is complete, your metaevaluation can be formative and provide feedback for changes.

Relevant Evaluation Standards

We consider the following evaluation standards, which are listed in their entirety in Appendix A, to be relevant to this chapter's content:

U1—Evaluator Credibility
U6—Meaningful Processes and Products
U7—Timely and Appropriate Communicating and Reporting
F1—Project Management
F2—Practical Procedures
F4—Resource Use
P2—Formal Agreements
P3—Human Rights and Respect
P5—Transparency and Disclosure
P7—Fiscal Responsibility
A1—Justified Conclusions and Decisions
A2—Valid Information
A3—Reliable Information
A5—Information Management
A6—Sound Designs and Analyses
A8—Communication and Reporting
E1—Evaluation Documentation
E2—Internal Metaevaluation
E3—External Metaevaluation

Case Studies

For this chapter, we recommend three interviews that illustrate different aspects of managing a project: *Evaluation in Action*, Chapters 6 (Rog), 11 (Ross Conner), and 13 (Allan Wallis and Victor Dukay)

In Chapter 6, Rog describes a national, multi site evaluation study of homeless families. She describes how she works closely with the funder and the sites, and how she seeks their input, and changes the evaluation, on critical issues. She also shows how she adapts the evaluation plan as the project proceeds and as they learn that homeless families differ from their assumptions. The journal source is Fitzpatrick, J. L., & Rog, D. J. (1999). The evaluation of the Homeless Families Program. A dialogue with Debra Rog. *American Journal of Evaluation, 20*, 562–575.

In Chapter 11, Conner describes how he conducts and manages a multi site evaluation of community health in Colorado using site evaluators at each location, visiting sites himself, and working from the beginning with the sponsor to plan the projects, provide assistance to sites, and move to evaluation. The journal source is Christie, C., & Conner, R. F. (2005). A conversation with Ross Conner: The Colorado Trust community-based collaborative evaluation. *American Journal of Evaluation, 26,* 369–377.

In Chapter 13, Wallis and Dukay describe how they put together an American evaluation team in preparation for their first visit to the orphanage they will evaluate in a small village in Africa. Dukay has developed this orphanage and now moves to the role of evaluator. They also put together a team of Tanzanian academics and townspeople to assist in the evaluation. This chapter is only available in *Evaluation in Action.*

Suggested Readings

Bell, J. B. (2004). Managing evaluation projects. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (2nd ed.). San Francisco: Jossey-Bass.

Compton, D. W., & Braizerman, M. (Eds.). (2009). *Managing program evaluation: Toward explicating a professional practice.* New Directions for Evaluation, 121. San Francisco: Jossey-Bass. Note: This issue is primarily about managing an evaluation unit in an organization and provides some excellent chapters on the issues as well as chapters by managers of evaluation units in organizations.

Perry, K. M. (2008). A reaction to and mental metaevaluation of the Experiential Learning Evaluation Project. *American Journal of Evaluation, 29*(3), 352–357.

Russ-Eft, D. F., & Preskill, H. S. (2001). *Evaluation in organizations: A systematic approach to enhancing learning, performance, and change*. Cambridge, MA: Perseus.

Stufflebeam, D. L. (2000). Lessons in contracting for evaluations. *American Journal of Evaluation, 21*, 293–314.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). Budgeting evaluations (Chapter 22) and Contracting evaluations (Chapter 23). *Evaluation theory, models, and applications.* San Francisco: Jossey-Bass.

This page intentionally left blank

Part IV

Practical Guidelines for Conducting and Using Evaluations

In Part Three, we provided guidelines for getting an evaluation started, including how to determine what should be evaluated, how to be certain the evaluation is focused on the right things, and how to plan the specifics of an evaluation. In Part Four, we focus on guidelines for actually conducting and reporting evaluations. Our focus in Chapters 15 and 16 is on the different choices evaluators make about data collection. The choices are presented in the order in which they occur. That is, first evaluators make choices about what design(s) to use to answer the specified evaluation questions in the context of the program being evaluated. With designs specified, they can begin to consider sampling strategies. Chapter 15 covers commonly used designs and sampling methods. Evaluators can then consider the appropriate means for collecting the information specified in the evaluation questions and the methods for analyzing and interpreting it. These methods are reviewed in Chapter 16. Our discussion includes both qualitative and quantitative designs and methods supporting our view that the choice is not based on a category of methods, but on the nature of the evaluation questions to be addressed and the context of the program. In most evaluation studies, a mix of qualitative and quantitative methods is used. In Chapter 17, we describe different ways to report the results to maximize the use or influence of the evaluation, and we discuss different types and frameworks for use.

None of the chapters in this section purport to treat completely their respective topics. A textbook could be devoted to each chapter (and, indeed, many texts do exist on several of those topics). Each chapter is intended only to introduce the topic and provide practical suggestions for how to use the material in an actual evaluation. References to more extensive discussions of each topic are provided for those who wish more detailed information.

This page intentionally left blank

15

Collecting Evaluative Information: Design, Sampling, and Cost Choices

Orienting Questions

1. What steps does one take to plan and carry out data collection for an evaluation?
2. How does using mixed methods and drawing from different paradigms increase the quality of an evaluation? How are mixed methods used most effectively?
3. How does the evaluator decide what to collect information on?
4. What are the purposes of each type of design? Under what circumstances would each be used?
5. When is sampling important in an evaluation? What steps would one take to select appropriate cases in a case study? To select a sampling strategy when the purpose is generalizing to a larger group?
6. How does cost–benefit differ from cost-effectiveness?

> The government is very keen on amassing statistics. They collect them, add them, raise them to the nth power, take the cube root, and prepare wonderful diagrams. But you must never forget that every one of these figures comes in the first instance from the village watchman, who just puts down what he pleases (Sir Josiah Stamp, as quoted by Light & Smith, 1970).

The collection of information is fundamental to evaluation. Although a policy maker or citizen may joke about "data-free" or "fact-free" evaluations, no reputable evaluator would presume to make evaluative judgments without first assembling a solid base of evidence.

Still, information is situational and changes with every evaluation. Similarly, the methods evaluators use to collect information also change to match the evaluation questions, the context for the program, and the information needs and values of stakeholders. Rarely is there one clear choice. Even the most experienced evaluator must make tough choices about sources, methods, and means for collecting information. The evaluator must consider the constructs to be measured; the reliability, validity, feasibility, and propriety of possible methods for collecting information; and the cost and manageability of the plan; as well as the political advisability and acceptability of it to various stakeholders.

In most evaluation studies the essential steps to collecting information are:

1. Study the evaluation questions that have been developed in the evaluation plan and determine what information should be collected. What information is necessary to answer the questions? What are the constructs that must be assessed or described? The selection of data-collection methods is driven by the nature of the evaluation questions.

2. Consider the nature of the program to be evaluated and its context, which includes its clients, funders, supporters, and other stakeholders as well as the community in which it resides. The evaluation questions have defined the information needs for key stakeholders in reference to the program, but what types of evidence will be most credible to these stakeholders and this community? What designs and methods are most feasible?

With that foundation, evaluators should take the next steps:

1. Develop or select a design(s) for collecting the necessary information. What design or designs are most appropriate for answering each question? How should the design be adapted for the specific circumstances of the study?

2. If appropriate, consider sampling strategies. Will information be collected from everyone or every site, or are the number of people or sites sufficiently large that sampling might be cost-effective? Is there interest in sampling a particular subgroup? If so, what sampling strategy is most appropriate for your purposes?

3. Identify appropriate sources and methods for collecting the information and areas where multiple measures are necessary. Who has this information (source or sources)? What is the most appropriate method for collecting it?

4. Develop procedures for collecting the information. Who will collect it? When? How? With what training and instructions?

5. Collect the information, using appropriate checks.

6. Analyze the information. Are statistical methods appropriate? If so, what statistical tests will be used? How should qualitative information be organized and interpreted? What software is needed?

7. Interpret the results and draw evaluative conclusions.

Entire books have been written on a single data-collection technique, instrument, or procedure. We could not hope to cover in comparable detail the many data-collection methods that evaluators have come to use and value. Instead, we devote two chapters to these topics. In this chapter we will discuss the steps evaluators take to answer questions 1–3 in the previous list. We will review issues concerning data collection, analysis, and interpretation in Chapter 16.

Using Mixed Methods

Evaluation Controversies over Methodology

From the 1970s through the early 1990s, evaluation was consumed by controversy over the use of qualitative or quantitative methods. The controversy over methods, while sometimes divisive, was useful for informing evaluators from different backgrounds and disciplines about the alternative approaches and measures available and for encouraging evaluators to consider multiple sources and methods. Evaluators address many different types of questions, from needs assessments to outcome and cost studies, and measure a wide array of concepts—from blood pressure to self-esteem, from computational skills to quality of life. Given the breadth of their tasks, evaluators must have a broad array of tools that encompasses both qualitative and quantitative methods.

Professional evaluators now agree that no one method or approach is always appropriate. In fact, one should not consider the method without first carefully contemplating the evaluation questions, the context and characteristics of the program, and the values and perspectives of various stakeholders. To make this point, let us quote two experts. Chelimsky (2007), in a volume on methods choice, observes:

> Selecting a method a priori without regard for the origins and specifics of the evaluation question clearly puts the cart before the horse. It is as if the Department of Defense were to choose a weapon system without regard for the kind of war being fought; the character, history, and technological advancement of the enemy; or the strategic and tactical givens of the military campaign. (p. 14)

Julnes and Rog, the editors of this volume, observe that each of the well-known evaluators who contribute chapters to the volume—evaluators from a variety of perspectives—agree on this issue. Good designs and methods do not stand alone. A good design choice is one that matches the evaluation question and the context.

Nevertheless, evaluators have continued to disagree vociferously on methodological issues. During the 1980s and early 1990s, the disagreements concerned qualitative versus quantitative measures. Today disagreements concern randomized clinical trials (RCTs). Evaluators, however, are not alone in their disagreements over method. Those studying what many consider the

"hard science" of medicine struggle with similar issues. Sackett and Wennberg (1997) observe:

> Lots of intellectual and emotional energy, ink, paper, and readers' precious time have been expended comparing, contrasting, attacking, and defending randomised control trials, outcomes research, qualitative research, and related research methods. This has mostly been a waste of time and effort, and most of the disputants, by focusing on methods rather than questions, have been arguing about the wrong things. Our thesis is short: the question being asked determines the appropriate research architecture, strategy, and tactics to be used. . . . (p. 1636, cited in Schwandt, 2007)

So, we move on to focus on the right things. We do not organize our methods by qualitative or quantitative categories, but by the different choices evaluators (and researchers) make—choices about the design of the study, samples to be selected, means for collecting data or information, and methods for analyzing and interpreting it. Our recommendations are not for a type of method, but for choices that make sense for the evaluation questions to be answered and the context of the study. Typically, these are mixed methods because few questions can be answered by only one strategy. However, if a question and its context lend themselves to one strategy, evaluators can certainly make that choice.[1]

Today's current disputes concern the attention or priorities being given to randomized clinical trials in some federal agencies in the United States and in other countries, particularly in the European Union. (See discussion of current trends in Chapter 2.) Yet, for all the differences of opinion concerning this new methodological controversy, the policies on which the arguments are based are changing and, in fact, provide more flexibility than some have implied. (see Lipsey, 2007). We will discuss RCTs later in the chapter, but remember that evaluators' goals are to use social science and its multiple methods to inform policy and improve decisions. Greene, a qualitatively-oriented evaluator, and Henry, a quantitatively-oriented evaluator, close their discussion of the qualitative-quantitative debate in evaluation in the *Encyclopedia of Evaluation* with this point:

> We, quantitative and qualitative evaluators alike, should unite in worry that the absence of evidence that meets some narrowly drawn standard will become a license for actions based entirely on ideology or the force of unconstrained rhetoric.

[1]Each of the well-known evaluators interviewed in *Evaluation in Action* used mixed methods. Some placed more emphasis on the quantitative data because such data were central to the questions they were answering and the types of evidence valued by their key stakeholders. (See, for example, interviews with James Riccio and Leonard Bickman.) But they also included interviews with clients and descriptive reviews of records to learn more about the program and its delivery. Others placed more emphasis on the qualitative data because they were central to the questions they were answering and the types of evidence valued by their key stakeholders. (See David Fetterman and Jennifer Greene.) But they, too, used quantitative data, surveys of program participants, to learn more about the programs they were evaluating.

> We should also unite in our commitment to enact our hard-won acceptance of multiple methods and multiple ways of knowing, to reclaim the conversation about the contributions of social science to social policies and programs and refocus it on substance and values rather than on method, and thereby to redirect our collective evaluation expertise and energy in the service of democratic social betterment and social justice (Greene & Henry, 2005, p. 350).

A Definition and Discussion of Mixed Methods

John Creswell (2009), writing about research design, defines mixed methods research as

> An approach to inquiry that combines or associates both qualitative and quantitative forms. It involves philosophical assumptions, the use of qualitative and quantitative approaches, and the mixing of both approaches in a study. Thus, it is more than simply collecting and analyzing both kinds of data; it also involves the use of both approaches in tandem so that the overall strength of a study is greater than either qualitative or quantitative research. (p. 4)

Jennifer Greene, while acknowledging differences in philosophical assumptions among evaluators, defines mixed methods evaluation more simply, "Mixed method evaluation involves the planned use of two or more different kinds of empirical designs or data gathering and analysis tools in the same study or project" (2005, p. 255). Having moved past the qualitative-quantitative debate, she, like us, emphasizes that evaluators can choose different methods at the key methodological stages of design, data gathering, and analysis. She notes that with the growth in use of mixed methods in evaluation, such that it has become a routine practice, paradigmatic or epistemological differences have become less important.

Moving beyond definition, she notes that mixed methods can be used for a variety of purposes, including:

- Triangulation
- Development
- Complementarity
- Initiation
- Value diversity (Greene, 2005, p. 255).

Triangulation is an older form of mixed methods, originated by quantitative researchers who used different methods to collect data on the same construct. (Picture the construct as the area inside the triangle and the sides as the different measures.) Their purpose was to increase the validity or accuracy of their measure of the construct as a whole. Putting together the results of three different measures, all intended to address the same abstract construct but in different ways, would increase both the validity of the study's measures of the construct as a whole and one's understanding of the construct. Two different paper-and-pencil tests would,

in all likelihood, be insufficient in breadth to provide a full picture of the construct of interest. Similarly, individual interviews and focus groups with the same individuals over a short period of time would provide some useful information, but the methods fail to differ sufficiently to constitute a complete assessment of the phenomenon of interest. Instead, evaluators might use a combination of a test and interviews to triangulate their measures and, thus, increase the validity of their measurement of the construct.

Greene and Carracelli (1997) studied 57 evaluations that used mixed methods and identified other purposes for using them. They found that mixed methods were frequently used to help develop other subsequent measures (development purposes). For example, evaluators might use a focus group or a series of individual interviews to help identify factors to be studied with a larger group using a survey. Alternatively, focus groups or interviews might follow a large survey to learn more about surprising and/or ambiguous responses to the survey. In addition to using mixed methods to increase validity or to develop future measures, Greene and Carracelli found that mixed methods were often used to gain a fuller understanding or picture of the construct of interest (complementary purposes). Methods with different biases may still be selected, but not with the hope of results converging and increasing validity. Instead, the hope is for somewhat different results that, when combined across methods, will provide a fuller picture of the abstract constructs we tend to examine in evaluation. Finally, they found that mixed methods were used to spark new ideas and thinking (initiation purposes). This may occur when measures that the evaluator thought would yield similar results instead diverge. Evaluators and others will then reflect, consider their models and others, and use mixed methods to collect new data to help explain the surprising divergence.

Mixed methods may also be used to represent different values. Some stakeholders may find certain types of measures or evidence more credible than other types. In one evaluation, stakeholders' values may differ and evaluators may choose to use different methods to meet these different needs or values.

Thus, mixed methods serve many purposes. They can be used to improve validity or understanding or diversity. Mixed methods also involve many different choices—choices concerning design, data collection, and analysis. In the next sections, we will describe different types of designs that might be combined or adapted in a study. In some cases, the evaluation questions may not require mixed methods; the questions may be readily answered by one type of design. In other cases, though, more than one design will be needed. When considering using multiple designs in a study, consider the following: (a) Will the designs be implemented concurrently or sequentially? Do the findings of one design inform the characteristics of another so that sequential timing is best? Or will concurrent timing help you learn more about the phenomenon being examined as it exists, not after it may have changed? (b) If the timing is sequential, the evaluators should be in an iterative and flexible mode, stopping to contemplate what has been learned at each stage and to redesign for the next stage.

Designs for Collecting Descriptive and Causal Information

Selecting the design or designs to be used in the evaluation is one of the most important decisions evaluators make. The value of different types of designs, particularly those used to address cause-and-effect questions, remains a bone of contention in evaluation, and we will examine those issues. But, just as with other methodological decisions, a good design is one that addresses the evaluation question's intent and is appropriate for the context of the program and the values of the stakeholders. In this section, we will describe different types of designs, the types of questions they are most likely to answer, and some important details or concerns in implementing each design. Table 15.1 provides a preview of the designs we will discuss.

Descriptive Designs

Although causal designs tend to get much of the attention among both evaluators and policymakers, descriptive designs are the most common in evaluation. Why? Because many evaluation questions are descriptive. Some examples include:

- Questions about needs or priorities: In which subjects have our students improved the most compared with previous years? What types of students perform most poorly? What courses or instructional strategies have they experienced? Answers to such questions can be used for program planning.

- Questions describing those receiving program services: What types of students participated in the program? To what extent were they like or unlike the target audience the program was designed for? Did their differences affect the delivery or potential outcomes of the program? (Example: The unemployment program is designed to provide job-seeking and interviewing skills. Those attending have these skills, but do not have skills that are marketable in the current slumped economy, so the designed program is inappropriate for this audience's needs. Another example: The after-school tutoring program is designed for kids who may get into trouble after school and who need tutoring. But, those who attend are students who perform well and are not at risk of poor performance or gang involvement.)

- Questions concerning program delivery: Were key pieces of the program delivered as planned? With the intended quality and duration? (Example: Many programs are delivered as planned, but others are not due to lack of training, lack of qualified personnel, differences among staff over the aims and activities of the program, insufficient resources or time, etc.) These differences in implementation should be described, for a variety of reasons, including to learn what did happen, to determine whether the model was delivered accurately so that evaluations of outcomes would actually be a test of the model, and to identify successful adaptations that might be used in subsequent deliveries of the program.

TABLE 15.1 ***Characteristics of Commonly Used Designs***

Designs	*Characteristic*	*Purpose*	*Sample Questions*
Descriptive Designs			
Case Study	Focus on a case Multiple measures Qualitative emphasis	In-depth description, understanding	Why are students dropping out of school? How do medical personnel and patients interact in the new lengthened sessions?
Cross-Sectional	Quantitative, survey	Picture of a group's behaviors, attitudes, characteristics, beliefs	How did the staff deliver the program? What problems did they encounter?
Time-Series	Examine trend Use existing data	To look for changes over time	Are the proportion of students using mass transportation to get to school increasing?
Causal designs			
Posttest-only Pre-post	Random assignment	Examine cause-and-effect relationships between the programs and outcomes	Did the students in the study lab improve their study skills? Their grades?
Quasi-experimental			
Interrupted Time-Series	Examining a trend for change Need existing data	Same	Did tardiness decrease with the high school's later start time?
Comparison Group	Compare two groups Select like groups	Same	Did student achievement improve more in schools that implemented teacher governance?
Case Study	Evaluate progress on logic model and explore contributions to program success or failure	Explain cause-and-effect relationships	What effects did the comprehensive educational reform program have and how were they achieved?

• Questions concerning program outputs, outcomes, or impacts: How much did students in the fitness program exercise during class? Outside of class? Did they achieve the desired weight loss or fitness goals? (Note this question is not causal. It is simply describing the participants at the conclusion of the program. If the goals are obtained, one cannot necessarily conclude that the changes are due to the program. But often outcomes are described first, before undertaking a more complex causal design. If the descriptive study shows very poor performance on the intended outcomes, the evaluator and other stakeholders may decide not to continue with a causal design.)

A Descriptive Study of Import. To illustrate the importance, and the sophistication, of descriptive designs to evaluation and decision making, consider a recent study reported in the press. The large, national study conducted by Westat, a national evaluation and research firm, was mandated by Congress and coordinated by the U.S. Department of Health and Human Services, and was the Fourth National Incidence Study of Child Abuse and Neglect (Sedlak, Mettenburg, Basena, Petta, McPherson, Greene, & Li, 2010). It was designed to determine the incidence of child abuse and neglect in the United States and the changes that had occurred since the previous study in 1993. The study made the front page of the *Denver Post* (Crary, February 4, 2010) because of its dramatic findings: They found that sexual abuse of children had decreased by 38%, emotional abuse by 27%, and physical abuse by 15% over the 12-year period from the last study to the current one (1993 to 2005–2006). Data on child abuse and neglect were collected from 10,791 "sentinels," as they were called in the study—people who might know of abuse, including people working in child welfare, police officers, teachers, health care workers, and day care employees. Data were collected from a nationally representative sample of 122 counties across the nation. The study, its sampling, selection of sources, and nature of the data collected were complex. As a descriptive study, replicating the earlier methods from 1993, it finds a sizable decline in child abuse—a positive finding. The study also provides information for improvement in child abuse policy, such as recommendations to assist in identifying abused children. The author, Andrea Sedlak, notes she is pleased with the decline, but "concerned that more than half of the cases were not investigated by child protective . . . There's still a lot of material here saying the system has a long way to go" (Crary, 2010, p. 9). We describe this study to illustrate the important kinds of information descriptive studies can provide.[2]

Now we will describe the three most commonly used descriptive designs: case studies, cross-sectional designs, and time-series designs.

Case Studies

The case study is one of the most frequently used designs in evaluation. It draws heavily on qualitative methods, but can employ both qualitative and quantitative methods. Although we are introducing it under descriptive designs, case studies can also serve causal purposes and we will discuss its purpose in that section as well. Two Roberts, Robert Yin (2009) and Robert Stake (1995), each of whom have written textbooks on the case study method, serve as examples of how

[2]The study also illustrates the tendency of the media to confuse description with causality, as much of the newspaper report, including the headline and interviews with other experts, concerns the "causes" for this decline. The headline is "Child-Abuse Crackdown Makes Huge Dent in Cases." The text of the article is ambiguous on the "cause," stating, "Experts hailed the findings as proof that crackdowns and public awareness campaigns had made headway" but later discussed other potential causes, including the increase in arrests and prison sentences (the crackdowns), changes in norms, and mediation for sexual abusers.

evaluators can approach case studies from different perspectives. Stake relies on a more interpretivist approach with a strong qualitative emphasis. (See Chapter 8.) His interest is in describing and understanding an individual case, while Yin's emphasis is on extending from that case to build knowledge or theory. Yin approaches case studies from a post positivist tradition, mixing qualitative and quantitative methods to achieve the three purposes he articulates for case studies: description, exploration, and explanation.

Case studies are particularly useful when the purpose of the evaluation is to describe something—a case—in depth. Very often, evaluations are concerned with exploring the "how's" and "why's" of a program or policy. How did a program achieve certain outcomes in clients? How was the program adapted in the field? Why did parents choose a charter school? Why do good teachers leave our district? How do students who transfer into college make connections? Case studies can be an excellent method for answering these types of questions because they encourage a deeper exploration of the issues, recognizing that there are many different perspectives on each.

The focus of a case study is on the case itself. Such an approach is particularly appropriate in an evaluation when there is a need to provide in-depth information about the unit or case at hand and not so much need to generalize to a larger population. Because evaluation is typically intended to be situation-specific, a case study design is useful because it uncovers the unique attributes of an individual case. Generalization to other settings or other times, which is often the focus of research, is not the goal in evaluation. Stake (1995) writes: "The real business of case study is particularization, not generalization. We take a particular case and come to know it well, not primarily as to how it is different from others but what it is, what it does" (p. 8).

Case studies are characterized by three features:

1. A focus on a selected case or cases
2. A desire for in-depth understanding of an issue
3. Collection of data in many different ways, but with a focus on qualitative methods such as observations, interviews, and the study of existing documents

Selecting the Case. The first challenging step can be selecting the case. Stake (2000a) observes that a case can be as broad or narrow as desired: "A case may be simple or complex. It may be a child, or a classroom of children, or an incident such as a mobilization of professionals to study a childhood condition" (p. 436). A case might consist of one large unit (a city, a school district, a hospital) or a few small units (classrooms, wards), or individuals themselves (students, clients, managers, providers). Cases might be selected because they are considered typical or because they are unusual. Unusual cases may be selected because they are particularly successful or because of their failures or for other reasons. The choice is up to the evaluator and others involved in planning the study, but the rationale for the choices should be clearly articulated. In some cases, the evaluator does not

make a choice: the case is the program to be evaluated, and it is sufficiently small that the entire case is studied.

Some examples may prove useful here. Consider a multisite program in which many sites are struggling to implement a program or to achieve certain outcomes, but some have succeeded under adverse circumstances. An evaluator might conduct a case study of one or more of the sites that have succeeded to explore the factors that facilitated that success.[3] Best-practice studies—identifying units or organizations that do well at something and describing them intensively—are broader examples of instances in which the case selected is not intended to be typical. One can see such choices in case studies of schools that have received high scores on state standards tests with at-risk kids. Other evaluations may select cases that are typical or may reflect the range of behaviors. An evaluation conducted by one of our authors included case studies of four female substance abusers to learn more about the difficulties they faced and the strategies they used in their transition to the community after staying in a residential substance-abuse facility. In the first example, the case is an organization—a school—that succeeded. In the second example, the cases are individuals who were selected to represent some typical outcomes.

Collecting Information. Unlike quantitative designs, a case study does not have a clearly delineated method. In fact, people even differ in how they label case studies. Some may call it a design; others, a method; others, an approach (Patton, 2001). Stake has written: "Perhaps the simplest rule for method in qualitative case work is this: Place the best brains available into the thick of what is going on" (p. 242). He writes that the person doing the case study should make use of her or his observational and reflective skills to obtain a greater understanding of the case at hand.

While the methods used in case studies are not as precise as for some other types of designs, case study methods can be characterized by the use of multiple methods and a greater emphasis on qualitative methods such as observations, interviews, and the study of documents. As noted, the goal is for depth in the description, for understanding. The methods of the case study may be selected or adapted as the evaluator achieves a better understanding of the case. That is, the design is iterative, responsive to the case and the circumstances at hand. It is adaptive and continues to adapt until evaluators believe they have a good understanding of the case.

Although the focus is on qualitative methods to attain that depth of description and understanding, case studies do make use of quantitative methods as well. Surveys, statistical analysis of existing data, and the like can be used to supplement what the evaluator has learned through observations and interviews. Methods may be selected as the study moves along and the evaluator identifies directions

[3]Brinkerhoff's Success Case Method (Brinkerhoff, 2003) compares cases that have succeeded with those that have failed, typically individual program participants, to identify possible reasons for success.

of uncertainty, areas where more information or understanding is needed. The evaluator can then select the method that is appropriate to examine an issue at each stage of the design.

Reporting a Case Study. Writing the results is an integral part of the case study, as results should be conveyed in a way that "focuses the reader's attention and illuminates meanings" (Guba & Lincoln, 1981, p. 376). Guba and Lincoln describe the case study as "holistic and lifelike. It presents a picture credible to the actual participants in a setting, and it can easily be cast into the 'natural language' of the involved audiences" (1981, p. 376). Such case studies can lead to greater use because the report is both more easily understood and more compelling than the typical report. Yin (2009) proposes six structures for writing a case study that can be selected and adapted for the purposes of an evaluation: linear–analytic (akin to a traditional research format); comparative (comparing different cases); chronological (telling the story in time sequence); theory-building (contrasting findings with a theory or model); suspense structures (beginning with an outcome and proceeding to explain how it was attained); and, finally, unsequenced structures (often used for descriptive evaluations). Lincoln and Guba (1985) discuss the case reporting format, and Hebert (1986) provides an example of the case study in the qualitative style. Stake (1995) presents a very useful example of a case study, with his own commentary, as the last chapter in his book. For more information on case studies, see Stake (1995) and Yin (2009).

Case studies are often used for descriptive purposes when the desire is to examine an issue from many different perspectives. Two other designs, more quantitative in their approach, can form a part of a case study or can stand alone as designs to answer descriptive questions: the cross-sectional design and the time-series design. Unlike the qualitative case study design, these designs do not provide in-depth descriptions. They can be fairly simple designs but are used frequently to answer rather straightforward questions.

Cross-Sectional Designs

The cross-sectional design is intended to show a snapshot in time of how people think, believe, or behave. A political poll is a common example of such a design. The attitudes collected are considered true for that particular point in time. Subsequent polls show changes in attitudes. Cross-sectional designs typically make use of a survey approach to collect information on the attitudes, behaviors, opinions, or lives of various groups, either total populations or subgroups sampled from those populations. The purpose of the cross-sectional design is both to describe trends across all groups and to identify any differences among the subgroups. (Recall how political polls, as reported in the media, help us learn both who might win—the overall trend—and which subgroups favor which candidates. The term "cross section" comes from examining subgroups.) A common example of using a cross-sectional design to obtain evaluative information is organizational surveys of clients or employees. Most organizations occasionally, or routinely,

survey their clients to obtain feedback. Annual surveys of parents are mandated in most schools. Hospitals survey former patients. However, such routine practices often don't make full use of the design because they fail to consider the evaluation questions that could be explored by such routine data collection or to identify subgroups that might be of particular interest for analysis.[4]

A cross-sectional design might be used to answer any one of the following questions: A principal asks, "What do parents think of our school? What do they see as the strengths and weaknesses of the school environment, facilities, curriculum, personnel? Do parents differ in their opinions based on the grade level of their child? Their child's performance? Their ethnicity? The parents' education and expectations?" The director of an outpatient unit of a mental health center asks, "How do our clients hear about us? What are their expectations about mental health treatment? What problems typically prompt their first visit? Do these opinions differ by the age, income, education, or ethnicity of clients? By the nature of their presenting problem?" These questions might be posed in the context of a needs assessment or a formative evaluation. At this initial stage, the primary interest is in identifying problems or priorities. Further evaluation may move into a case study mode to explore the viability of solutions to problems discovered through the cross-sectional design.

Time-Series Designs

A time-series design is intended to demonstrate trends or changes over time. Unlike the interrupted time-series design, its purpose is not to examine the impact of an intervention, but simply to explore and describe changes in the construct of interest, as in the study described previously on the incidence of child abuse. The results of a time-series design can be very useful at the beginning stage of a case study if the evaluator explores with stakeholders their interpretations of the ups and downs exhibited in the results. Their perspectives may point the way to the next steps in data collection.

As with the cross-sectional design, the questions to be answered with a time-series design are relatively simple and straightforward. A health administrator might ask, "Is the number of premature births in our hospital declining?" A high school principal may ask, "Is the proportion of our student body needing ESL classes increasing or decreasing?" A police chief might ask, "What is the trend of juvenile crime in our city? Which juvenile crimes are increasing? Which are decreasing? Which are remaining stable? How do these trends compare with the number of juveniles in our population? Will the number of juveniles remain the same in the next decade?" The latter includes a number of

[4]One of our authors has worked with schools to insert a few questions into their annual surveys of parents to collect data for decision-making purposes. For example, in times of tight resources, a routine parent survey was revised to gather opinions on areas that might be cut. In this way, the principal or superintendent learns from a more representative group of parents, not just the more vocal ones.

different questions that will help the chief and her staff in planning, but all would be addressed through simple time-series designs.

Time-series designs generally make use of existing information to obtain enough observations over time. Key decisions involve the time ranges to use (daily, weekly, quarterly, semiannually, annually) and the number of data-collection points to obtain. As evaluators collect information from points increasingly further back in time, they must make sure the data-collection methods themselves have not changed. Changes in the manner of data-collection or definition of terms (What is a juvenile? What is a felony? Which crimes are recorded? What is an ESL class? What is a premature birth?) can make it appear that there is a change in the phenomenon being studied when, in fact, the change is due to the manner in which data are collected or categorized. (See O'Sullivan, Rassel, and Berner [2003] for more information on time-series and cross-sectional designs.)

Causal Designs

Causal designs are, of course, intended to answer causal evaluative questions. These are questions of the sort: Did X program or policy cause Y outcomes? In such cases, stakeholders do not just want to know if the program outcomes were at the desired level or whether achieving program outcomes is associated with attending the program. They want to know if the program itself caused those outcomes to change. (Whether the change is to the desired degree will be determined by interpreting the data and comparing it with standards set during the planning stage of the evaluation. Causal designs focus simply on whether the observed changes can be attributed to the program.)

Clarifying Stakeholders' Expectations and Understandings. Stakeholders are often interested in outcomes, but they may or may not be interested in establishing causality. They may not, in fact, understand the difference. As shown in the study of trends in child abuse, media and others who are uninformed about research methods are much more willing to see desired changes as being due to a program or policy even if the design was not intended to establish a causal link between the program and the reported outcomes. In some cases, for political reasons, stakeholders would prefer not to examine causality. If the desired goals are achieved, many, particularly supporters of the program, would rather stop there and assume program success. Therefore, evaluators must carefully examine and discuss stakeholders' interests in outcomes to learn more about what they want to know and how they hope to use the information. As the *Standards* and Guiding Principles attest, evaluators are obligated to share with stakeholders the limitations of their methods or misconceptions that might arise. Specifically, Guiding Principle A2 indicates:

> Evaluators should explore with the client the shortcomings and strengths both of the various evaluation questions it might be productive to ask and the various approaches that might be used for answering those questions.

Guiding Principle A3 states:

> Evaluators should communicate their methods and approaches accurately and in sufficient detail to allow others to understand, interpret and critique their work. They should make clear the limitations of an evaluation and its results. Evaluators should discuss in a contextually appropriate way those values, assumptions, theories, methods, results, and analyses significantly affecting the interpretation of the evaluative findings. These statements apply to all aspects of the evaluation, from its initial conceptualization to the eventual use of findings.

Because of the public and other stakeholders' lack of knowledge about ways to determine causal relationships, these principles are much more pertinent to causal questions than to descriptive ones, although misconceptions or misunderstandings can occur with descriptive designs as well. Nevertheless, evaluators should discuss causal questions carefully with their clients and key stakeholders to learn more about their expectations and to educate them to the different types of information the evaluation could provide about outcomes. The client should understand the differences among and the implications of describing outcomes, establishing a causal linkage between the program and outcomes, and determining if the outcomes achieved are sufficient. These different purposes call for designs that differ in cost and complexity and have ramifications for flexibility in program operations.

Incidence and Importance of Causal Designs. Anyone who reads the newspaper is familiar with experimental designs, even if the person doesn't know the appropriate labels. Studies on the effectiveness of various drugs or medical treatments reported in the news frequently contrast findings from the group that received the new drug or treatment with those who received a placebo or an older, established treatment. Patients are randomly assigned to the groups. Health-care providers and patients are *blind* to the group they are in. Treatments are delivered, data are collected and analyzed, and conclusions are drawn. Such designs are commonplace in the medical literature and we, as health consumers, expect such rigorous, FDA-approved studies.[5]

If such designs are considered appropriate, even desirable, when our health is at stake, why not when the issues are learning, job training, counseling, child care, or the air we breathe? Because applying the medical model without thought to social policy issues is a mistake. Misuses of experimental designs have been legitimately criticized (House, 1990; Johnson et al., 2008; Schwandt, 2005). Rejecting such designs out of hand as infeasible, unethical, or uninformative, however,

[5]We are, of course, simplifying here. As we know from the scandal involving Vioxx—a drug purportedly tested carefully under FDA review and widely used for arthritis pain—such studies can misinform. The drug was withdrawn from the market in 2004 because it increased the incidence of cardiovascular events. The drug led to more than 27,000 heart attacks in patients between 1999 and 2003. So, even our supposedly straightforward experiments to test pharmaceuticals are not free from error or bias.

can needlessly limit evaluators' methodological choices. The medical model—and experimental designs—have limitations for studying human service programs because human behavior and "treatments" for human behavior are often much more complex and variable than a drug or surgical intervention. Evaluating the complex program theory or logic model of a substance-abuse treatment program or a reading comprehension program is more difficult than assessing the impact of a drug, because the implementation varies and the impact takes longer. A colleague of ours once remarked disparagingly, "This [studying public policy] isn't brain surgery!" To which one of us responded, "No, it's tougher. Human brains are relatively similar. Behavior is not. So, studying educational and social phenomena is much more difficult than brain surgery!"

Chelimsky (2007) wrote the following about her 14 years running the Program Evaluation and Methodology Division (PEMD) of the U.S. Government Accountability Office (GAO):

> Cause-and-effect questions were only a small minority of the questions posed to evaluators by policymakers in either branch. . . . This may be because other kinds of questions have had a higher political priority in the Congress or in various administrations than cause-and-effect ones did, or because the governmental or organizational context of the program or policy to be studied made the experimental design seem infeasible or inadvisable to practitioners and evaluators, or because policymakers are always and everywhere economical in their thirst for knowledge about program effectiveness, especially in areas where it could conflict with earlier legislation or administration imperatives. (p. 21)

Others (Datta, 2007) have found that use of causal designs varies with the culture of the agency rather than with the intent of the evaluative question. Some health-related agencies, such as the National Institutes of Health (NIH) and the Substance Abuse and Mental Health Services Administration (SAMHSA), prefer experimental designs, while the U.S. Department of Justice, the World Bank, and the U.S. Agency for International Development (USAID) make extensive use of case studies for causal purposes. (The Justice Department follows Yin's work on explanatory case studies.) The National Science Foundation supports a variety of designs as long as they are appropriate for the question being studied, and the Bureau of Indian Affairs (BIA) prefers ethnographic approaches that are more compatible with their culture. Although Mark and Henry (2006) have argued that policymakers prefer experimental designs to address causal questions, the evidence is mixed. Even when the key client is a policymaker, evaluators need to carefully assess their clients' expectations and information needs.

Even though causal designs are not used frequently, they receive much attention because they can address quite important questions about the ultimate merit and worth of programs and whether they deserve funding and continuation. As evaluators obligated to consider the public interest and the public good (see AEA Guiding Principle E), we must recommend causal, experimental designs when the questions and context merit it. Let us now examine a few of those designs so the reader can learn which might be most appropriate in a given context.

Experimental Designs

Experimental designs, or randomized control trials (RCTs), if feasible, are preferable to quasi-experimental designs in that they can counter more threats to the internal validity of the study. This means that, if the results from an experimental design show that those receiving the new program or treatment improved more than those receiving something else, the evaluators and audiences for the study can have more confidence that the improvement was due to the program and not to some other factor than would be the case if another type of design were used. Experimental designs include pre-post and posttest-only designs. Each of these designs involves randomly assigning program participants to a group. Through random assignment of a sufficient number of people to each group, experimental designs maximize the probability that the groups are equal on the factors that could influence their response to the program, such as individual characteristics and attitudes, past history, things going on in their lives currently, and so on. Although individuals in the groups are not equal, the groups as a whole are viewed as equivalent. Thus, differences that emerge between the groups on measures taken after the program can be more validly attributed to the program itself rather than to other causes.

Posttest-Only Designs. These are the least complicated of the experimental designs and require simply the following. First, one must decide what comparisons are desired and meaningful. For example, will the program of interest—in this case, an experimental educational program called X—be compared with an alternate educational program, Y, that attempts to achieve in some alternate way the same goals as those sought for Program X? Or will students in group X be compared only with those in another group that have received no similar or alternate program with goals similar to X? In other words, evaluators must consider what those who are *not* receiving the new program or intervention will receive. Typically, the control group receives whatever represents the choice for policymakers: Is the choice to be made between the new program or nothing, or is it between the new program and an existing one?

Second, steps must be taken to randomly assign those who will participate in the study to the various treatment groups.

Next, evaluators or researchers must monitor the treatment to ensure that those in one group are not influenced by the treatment received by those in the other, and that the program is delivered as planned.

Finally, information must be collected after the program ends (the posttest) to determine whether differences occurred.

The name of the design, posttest-only, does not dictate the measure to be used. Posttreatment measures can be surveys, interviews, observations, tests, or any other measures deemed appropriate for getting a full picture of the outcomes. The term"posttest-only" refers only to the time at which information will be collected. Multiple measures can be used if the constructs to be measured are complex or if there are several desired outcomes. No pretest information is collected with

the posttest-only design, because it is assumed that the two groups are equivalent due to the random assignment of individuals or units (offices, schools, classrooms) to the programs or treatments.

Pre-Post Designs. This design is used when a pretreatment measure can supply useful information. For example, if the groups are small, there may be concern about their equivalence. A pretest can help confirm their equivalence, though only on the measures collected. If there is concern that many participants may drop out of the program, and that the participants who are available for the posttest may no longer constitute equivalent groups, pretest scores can be used to examine differences in the remaining two groups. For example, dropouts would be a legitimate concern in evaluating a six-month training program for the hard-core unemployed, but would probably not be a concern in a month-long program for fourth graders at a school with stable enrollment. Thus, pretests can provide useful information with small groups or groups in which dropout rates may be high.

Many use pretests as benchmarks to report the change that has occurred in those participating in the program from before the program to its conclusion. These reports are often appealing to stakeholders; however, pre-post comparisons can be misleading because the change from pre to post can be due to the program and/or other factors in the participants' lives, such as natural changes that occur with the passage of time, other learning, and intervening events. Instead, the post measure of the comparison group is generally the more appropriate comparison, because that measure better represents what the treatment group would have been like at that point in time if they had not received the new curriculum or program. In other words, the comparison group experienced the same other factors and passage of time that contributed to change from the pre to the post measure in the group receiving the new program; they simply did not experience the program. So, the difference between the comparison group and the treatment group more clearly demonstrates the effect of the program than does a comparison of the change from the pre to post measures.

Feasibility and Propriety. Experimental designs require a good amount of experimenter control to randomly assign participants to groups and to avoid contamination between the groups—participants in one group being affected by the other. In many cases, random assignment is simply not feasible because the program, or more likely the policy, must be applied to everyone. For example, highway speed limits cannot be randomly assigned to study the effect of speed on highway safety. Clean air policies such as restrictions on wood-burning fires cannot be randomly assigned to households to examine their effect on air pollution. Students cannot be randomly assigned to charter schools to study the effects of charters or choice. In fact, selection is a major threat to validity in studying the effects of charter schools or choice on learning. Why? Because those parents who explore schools other than their child's neighborhood school, and perhaps select

a different one, are more interested in their child's education, have more time, or are more comfortable and confident making choices in the education arena than those who do not. Therefore, their children may score higher than their counterparts who stayed in the neighborhood public school for reasons other than the education provided by charter schools.[6]

Difficulties in using random assignment to assign participants to the program or to an alternative condition can limit the feasibility of using an experimental design, but other difficulties exist as well. Both the experimental treatment and the services received by those in the control group must be carefully monitored to ensure that they are delivered as planned, thereby confirming that the program is really what is being tested. In long programs, dropouts inevitably occur, so the two groups are no longer equivalent. Can the evaluator be sure that those who dropped out of the treatment group were equivalent to those who dropped out of the control? Extensive pretesting can be used to help adjust statistically for differences across groups in those who leave the study to maintain the equivalence of those who remain in the study. However, all these controls are costly and complex and are often simply not feasible in a typical organization. Therefore, when considering using an experimental design, evaluators must ensure that those working in and with the program, and the organization that contains it, are aware of the implications.

Many argue against random assignment to treatments on an ethical basis. Such concerns can be legitimate. Often new programs have been carefully planned, have a firm theoretical foundation, and offer great promise to participants. However, we often fail to consider the ethical issues involved in failing to study the new curriculum, policy, or program thoroughly. Is it ethical to expose people to treatments or programs that may, in implementation, be less successful at achieving the goal than the currently accepted program? Is it ethical to raise the expectations of those in need and then dash them with an untried, untested program? In a time of declining public resources, is it ethical to continue expenditures on an untested approach when these resources could be used to effectively meet needs in proven ways? There are no easy answers to these questions.

The consequences of randomization must be considered carefully for each circumstance. What are the risks to each group? How much do we know about the new treatment? About the old? How long will the study period last? How will participants be given informed consent? How will their dignity be protected? Under what circumstances could the data collection be halted and the better treatment delivered to all?

It is not uncommon for a new program to be costly in terms of either materials or training of personnel to deliver the program. For these reasons, or others,

[6]Some studies of school choice have made use of experimental designs. In such studies, parents are randomly assigned to be eligible for vouchers or not. The design can then track differences in students' achievement and attribute any differences identified to eligibility for a voucher to choose a school. Still, the results cannot be generalized to students whose parents are not initially interested in a voucher program. Internal validity is strong due to the design, but external validity, or generalizabilty, is not established.

it often is not practical to deliver a new program or curriculum to all students or consumers. In other cases, a new program may be controversial. Some stakeholders, including staff who deliver the program, may disagree on its merits relative to other options or the present mode of delivery. In such circumstances, random assignment to the new treatment can be the most fair option.

Robert Boruch (2007), who advocates greater use of randomized clinical trials, suggests that randomized designs be considered when the answers to the following questions are in the affirmative:

1. Is the social problem [that the program addresses] serious?
2. Are purported solutions to the problem debatable? In other words, are other options being considered, or is this the only option?
3. Will randomized trials yield more defensible (less equivocal and unbiased) results than alternative approaches to estimating effects?
4. Will the results be used?
5. Will human rights be protected? (p. 56–57).

We will add the question: Has the program or policy been pilot-tested and fine-tuned so that it is working as expected? There is no reason to rush into examining the causal outcomes of programs that have not had the opportunity to work out the kinks, yet in today's outcomes mania some policymakers rush to judgment.

Often, the answer to at least one of Boruch's questions is no. The problem the program addresses may not be that serious. Or, there may be a lack of resources or the will to ultimately use the results because the program is too costly to implement broadly. But if the answer to each question is yes, Boruch argues that randomized trials should be used:

> When we launch programs to redress serious social problems in the United States and in the developing world, we owe it to the people we are theoretically serving to get it right. No less than those of us taking our daily dose of aspirin, they deserve assurance that the interventions they are subject to are effective in improving their life chances. (2007, p. 72)

Quasi-Experimental Designs

For many evaluations concerned with establishing effects, random assignment is neither feasible nor desirable. In such cases, a quasi-experimental design can be more appropriate. These designs do not involve random assignment, but can be useful in countering some explanations for change other than the program. The most commonly used quasi-experimental designs are the interrupted time-series design and the nonequivalent comparison group design.

Interrupted Time-Series Design. This design involves collecting data many times prior to the program and then many times after its introduction. It is used frequently when the intervention, or program, is a law or policy that must apply to

everyone in the district, city, state, or nation. New clean air standards cannot be randomly assigned to some households and not others. Changes in the laws for prosecuting juveniles cannot be applied to some juveniles and not others. However, for each of these "programs," information is routinely collected on phenomena of interest prior to and after the new laws or standards are imposed. Environmental agencies routinely collect data on air quality; juvenile justice agencies collect such data on juvenile crime. These existing data can be analyzed to assess program effects.

Theoretically, an interrupted time-series design can be used in many settings. In fact, its most frequent application is with existing data that have been collected routinely prior to the intervention. The value of the interrupted time-series study lies in the measures made prior to the intervention. These measures help demonstrate trends or typical variations that occurred before the program was initiated. By having many measures collected before the policy is implemented, the evaluator can examine the typical ups and downs of the construct being examined, whether it is high school graduation rates, cancer survival rates, housing costs, or drug arrests. Having established the usual trend, the data can be examined to determine if the trend changes after the introduction of the program or policy. If so, the change may be due to the program (or to some other phenomenon that occurred at the same time). If, instead, the post measures simply reflect the usual variation, the results would suggest that the policy or program had no effect. (See Figure 15.1 for sample time-series lines and draw your own conclusions about program effect. Note that in some cases, one would conclude that the program was effective if only the immediate pre-program and post-program measures were examined.)

Observe that establishing the pre-program trend is critical with this type of design. This trend can only be estimated by having many data points prior to the intervention. One pre measure is not sufficient, because it might be an aberration caused by measurement error, an exceptional group or time, or a combination of all three. In fact, new programs are often initiated in response to an extremely poor previous year. The next year may have resulted in improvement by chance because the prior year was atypical. Nevertheless, if stakeholders examine data from only the prior year and the first year of the program, the program will look successful, simply because whatever is being monitored is moving back to the norm (or regressing to the mean, for the statistically sophisticated!). A current mistake of this nature is the tendency for most states to report changes in scores on school standards based on comparisons only with the previous year. No prior trend is established and, therefore, it is impossible to determine whether the increase or decline is due to real changes in teaching methods, changes in the student population in the school, changes in staff, or measurement error. Yet, the general public and many elected officials believe these annual changes reflect real change in each and every case.

Even when used correctly with ample pre-intervention data, a drawback remains to the interrupted time-series design. A change in the trend line certainly may indicate a program or policy effect, but the change may have been caused by something else occurring at the same time. Let's say that we have introduced a program for new teachers to improve retention and that we use an interrupted time-series

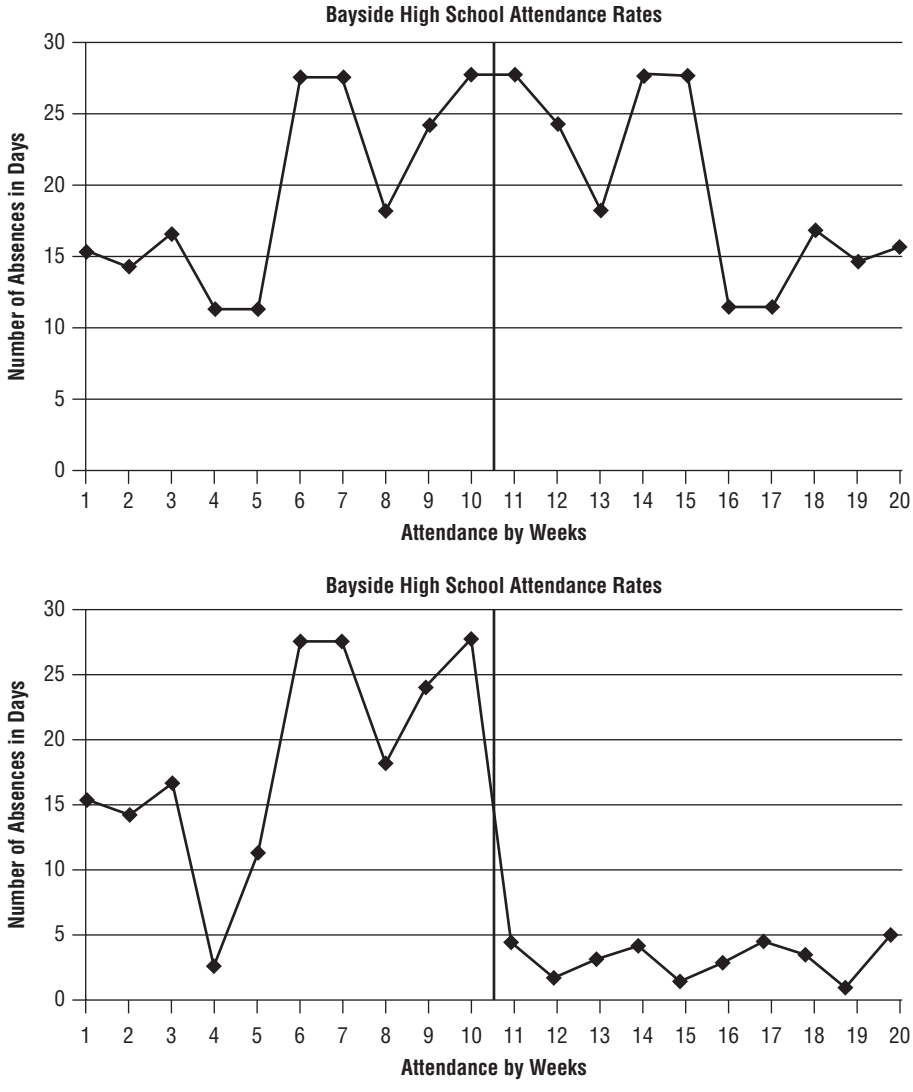


FIGURE 15.1 *Sample Time-Series Trends*

design to examine annual turnover of new teachers for the past 15 years. After two years, our chart shows a decline in turnover compared with the trend prior to the orientation program. This change could be due to our orientation, but what if the economy in our school district had taken a serious downturn during that time and jobs were scarce? The scarcity of jobs may have discouraged new teachers from changing jobs and, thus, improved our retention rate. In other words, when interrupted time-series designs are used to establish causality, evaluators, stakeholders, and other audiences should always consider whether something else that occurred

at the same time as the intervention could have brought about the change. A related problem: Very often in the public sector, we institute a package of several reforms to deal with a serious problem. This package may help us to address the problem comprehensively, but it hinders us in discovering which aspects of the package worked. The intervention or program should either be considered to include the entire package of reforms, or the other reforms must be viewed as alternative explanations for the apparent causal relationship between the single intervention and the outcome.[7]

One other caution: An interrupted time-series design is most appropriate with programs that expect a relatively quick change. If the change is gradual, the change in the trend line will be gradual and it will be more difficult to attribute any observed changes to the program. Of course, one can lengthen the time between points of data collection so that the trend line might show a more immediate effect, but the longer the time between points, the more likely it is that other factors may have caused the change.

In summary, consider an interrupted time-series design when the following conditions can be established:

- Random assignment is inappropriate or impractical
- Existing data are available that have consistently measured the construct of interest
- Quite a few data-collection points exist to establish a trend prior to the new program or policy
- Few, if any, other factors are anticipated to occur concurrently that could also change the construct
- The program or policy should have a relatively quick impact

Comparison Group or Nonequivalent Control Group Design. The comparison group design is similar to the experimental pre-post design, but participants or students are not randomly assigned to groups. Instead, we try to find an existing group that is very similar to the one that will receive the new program. The pretest is a more important component of this design than it is in the experimental designs because it helps us to examine similarities between the groups. Of course, the goal is to establish the equivalence of the groups, if only on the pre measure. (It would be wise to collect other descriptive data to compare the two groups and further explore their equivalence.) If intact groups in large organizations are being studied (e.g., offices, classrooms, schools, wards) and the program is short-lived, it may be relatively easy to find a comparable comparison group. However, if the organization

[7]An example: Colorado instituted a number of reforms to reduce car accidents, injuries, and deaths among new drivers who are the most likely age group to have a serious accident. Interventions included a longer trial period for obtaining a complete driver's license, restrictions on driving with young people other than siblings, more serious penalties during the trial phase, etc. Traffic accidents, injuries, and deaths among young drivers declined and the whole package may have been necessary and valuable. But, we would be unable to sort out which elements of the package of reforms had the most, or the entire, effect.

is small (a single elementary school with three classrooms per grade or a school district with two high schools), it is likely that the different units will have some significant differences. If the program is long, the groups may begin as relatively equal, but other differences may occur through the course of the program (e.g., different teachers or staff with different motivations, skills, and emphases) that could contribute to differences on the final measure. (See McCall, Ryan, and Green [1999] for a useful discussion of nonrandomized constructed comparison groups for evaluating age-related outcomes in children.)

Regression-Discontinuity Design. This design is particularly useful when eligibility for the program to be studied is determined by a person's scoring above or below a certain point on the eligibility criterion (e.g., high blood pressure or cholesterol levels). For example, patients may be eligible for a special weight-reduction program based on being at least 30 percent above standard weight guidelines for their height and gender. The design then compares outcomes for these patients with outcomes for people who were not eligible for the program, using regression methods. A "discontinuity" in the line, or a difference in the regression line, for the two groups suggests a program effect. This design can be useful when programs are limited to those most in need or most qualified, such as a program for highly gifted students, and eligibility is determined by a clearly defined cut point. See Trochim (1984) and Reichardt, Trochim, and Cappelleri (1995) for more information on this design.

Shadish, Cook, and Campbell (2002) provide more information on design in general and quasi-experimental designs in particular. One of the newer issues in experimental designs concerns the failure to adequately consider statistical power in planning designs. As a result, Type II errors—or failure to find significant differences between groups when such differences really exist—occur far more frequently than we are aware. Such errors can cause us to reject beneficial programs because we believe they make no difference when, in fact, small sample sizes and/or large group variability may have limited our ability to detect differences. Lipsey (1990) discusses methods for planning designs to avoid such problems.

Case Studies for Causal Purposes. Case studies are often used when the purpose of the evaluation is descriptive, but they can also be quite successful for examining outcome issues. Consider case studies you have read that illustrated how a program or curriculum was implemented—the factors that influenced program adoption, early stages of implementation, problems encountered, frustrations endured, surprises experienced, adaptations that were made, successes achieved, staff and participant reaction, environmental influences, and the like. Such studies give the reader a real understanding of the program and the many different ways it might be viewed. The voices and perspectives of many different stakeholders involved with the program are heard.

Robert Yin (2009) has long been an advocate of using case studies to *explain* the cause-and-effect links between programs and their outcomes. He observes that

experimental designs, because of the focus on control, struggle to explain those links in order to eliminate other sources of cause-and-effect or threats to internal validity. In so doing, an experiment, or a randomized clinical trial, is rarely like the true program because everything is controlled. If we want to see how a program really works and to explore and explain its working, a case study can be very useful. The strength of experiments, Yin notes, is in "establishing the efficacy of a treatment," but they "are limited in their ability to explain 'how' or 'why' the treatment necessarily worked, whereas case studies could investigate such issues" (2009, p. 16). The strength of case studies is not in control, but in attending to the context and the details. Again, Yin writes, "you would use the case study method because you wanted to understand a real-life phenomenon in depth, but such understanding encompassed important contextual conditions—because they were highly pertinent to your phenomenon of study" (2009, p. 18). In other words, Yin argues that case studies should be a preferred method to establishing causality in evaluations because their absence of control and their attention to context make their results more applicable to the real world. An experiment separates the program from contextual issues to attend to only a few variables—the purported cause (the program) and the effect (the desired outcomes). An explanatory case study examines the context, studying it to explain and understand the workings of the program.

How does a case study do this? When the purpose of a case study is explanatory, rather than descriptive or exploratory, the evaluator works with the literature and the people who developed the program to build what we might call a program theory and what Yin calls a chain of evidence. The program theory specifies the individual steps from start to finish for the changes the program hopes to bring about in those it is serving. The case study, then, collects information on these steps. Do they occur? How do they occur? Do changes take place at each stage? In all participants or only some? The case study, of course, uses a variety of methods to answer these questions or others that help to explain the program process and workings. Mark and Henry (2006), coming from a somewhat different perspective with less focus on case studies and more on causal methods, also discuss using program theory or theories of change (Weiss, 1995) or tests of mediation (Mark, 2003) to "assess how a program has its effects" (Mark & Henry, 2006, p. 319). Like Yin, their emphasis is on measuring these theories of change or mediators to explain the program. Acknowledging some of the drawbacks of experimental methods, they write:

> It is one thing to know, say, that a particular juvenile justice program works. It is another thing to know how it works. Does the program work because it avoids labeling, or because it enhances adolescents' self-esteem, or because it strengthens their ties to the community, or because it offers attractive activities that reduce opportunities for getting in trouble? (Mark & Henry, 2006, p. 319)

Now, an explanatory case study would not explore all those options, because it is focusing on evaluating a specific program that has a particular chain of reasoning or theory of change. The case study is collecting data on the steps in that

chain. However, evaluators' knowledge of other possible routes to change would prompt them, in a case study mode, to explore these other possible routes to success. The purpose is to document which route led to the desired outcomes. Of course, the most obvious route to study is the one proposed by the program developers. However, programs often succeed in different ways. The case study approach—putting evaluators close to the context and the program to understand rather than control—permits evaluators to explore different mechanisms for change and to document those that occur.

Yin provides an excellent example of using case study methods to evaluate comprehensive reforms or reforms that cover an entire community, city, or state and that have many different elements to the intervention. He argues that traditional experimental designs will not be adequate to study such reforms: Random assignment is not possible because the unit of analysis is quite large, a city or state. In addition, control of the intervention is difficult because it is multifaceted. Yet, such comprehensive reforms are becoming more frequent as government addresses more difficult problems. He cites examples of efforts toward systemic change and comprehensive reform in K–12 education, public health campaigns on prevention of drug abuse and HIV/AIDS, mental health, and community partnerships for social and economic development. Using as a first step the development of a logic model for the comprehensive reform, Yin describes how he uses the case study method to evaluate statewide comprehensive education reform in eight states. A previous evaluation had found little effect because it focused on a narrow definition of the comprehensive program. Yin, by studying many elements involved and affected by the comprehensive reform, found important effects beyond the efforts of the state education agencies. (See Yin and Davis [2007] and Yin and Kaftarian [1997].) He uses a similar method to study comprehensive reform and its effects in 27 urban school districts (Yin, Davis, & Schmidt, 2005).

Mixed Method Designs

The designs described thus far—the case study, experimental and quasi-experimental designs, and the cross-sectional and simple time-series designs—are all models or archetypes that the astute evaluator can use to craft a design or designs appropriate for the questions of interest in a specific evaluation. As we noted at the beginning of this discussion, there is no one design that is best for all settings. A good design is one that matches the evaluation questions developed during the planning phase, the context of the evaluation, and the information needs and values of stakeholders.

As evaluation has developed, evaluators have adapted these designs to different settings. Case studies can be complements to experimental designs to help explain any effects the experimental design detects. Cross-sectional designs or surveys can be used to complement experimental or quasi-experimental designs to learn more about how the program is delivered or, in the case of quasi-experimental designs, to overcome threats to internal validity or other explanations for the apparent causal relationship between the program and outcomes. Surveys of participants can be used to determine the extent to which participants in a comparison group design are equivalent and

whether their only differences, as groups, are in exposure to the program. Case studies, of course, use multiple methods and can incorporate time-series designs and cross-sectional designs within their broad purview.

So mixed methods are selected at the design stage. Evaluators make wise choices by considering these designs as archetypes and learning more about them. Then they must consider how to adapt or combine them to meet the demands of the individual evaluation questions in a study, the context in which the program takes place, and the values and information needs of diverse stakeholders. Meeting these demands is a challenging task and mixing or combining design elements is one way to accomplish it.

Sampling

Sampling is the method the evaluator will use to select the units (people, classrooms, schools, offices, counties, etc.) to study. If, in a statewide evaluation of immunization rates, we were to test only a portion of the children in the state, we would be sampling. If our sample were constructed using some randomized procedure, we might then be able to use the information collected from the sample to make inferences about immunization rates and patterns of the entire target population (in this case, all the children in the state). But if generalization is not our purpose, using such a mode of sampling would be inappropriate and a different sampling strategy should be used.

Sampling is not necessary in all evaluations. If the population of interest, or the group to which the results of the study will be extended, is small, it would be wise to collect information from the entire group. If the population is large, however, sampling can reduce the costs of collecting information. We would never attempt to conduct intensive interviews on 300 clients. Similarly, we would typically not survey 30,000 former graduates.

Sample Size

If sampling is required, the evaluator must first determine the appropriate sample size. The variability of the phenomenon to be examined and the desired degree of accuracy both affect the sample size. These, of course, require some judgments on the part of the evaluator, but if the phenomenon to be measured is quite variable and a relatively high degree of confidence is needed in the estimate, larger sample sizes will be required. For cross-sectional studies, many statisticians suggest trying to obtain at least 30 people for each *cell* or subgroup to be examined in the study. One can also obtain guidance by examining the sample sizes used in similar evaluations. Finally, statistics books provide details on power analyses to estimate desired sample sizes (Lipsey, 1990; O'Sullivan, Rassel, & Berner, 2003). Remember, too, that you are trying to estimate the final sample size you need, which means all those who respond to a survey or provide consent for you to obtain existing data. Because no survey or informed consent process receives a 100% response rate, you must also consider the loss of data from nonrespondents to estimate the sample you will draw.

Henry (1990) emphasizes that, in addition to the statistical techniques for estimating the desired sample size, evaluators should also consider the credibility of the size of the sample to significant stakeholders. He gives an example of a study in which evaluators selected a sample of 60 licensed homes for adults. After collecting detailed data, the evaluators found that the administrators using the study viewed the sample as too small. After further extensive data collection from 240 more homes, the final results were within 3% of the initial results with the smaller sample. The problem, therefore, was not with the original sample size, but the credibility of the size to the central audience for the study. Henry cautions, "Prior planning and attention to factors that may serve to undermine sample credibility may thwart undue attacks" (1990, p. 126). His advice and examples illustrate how involvement of significant stakeholders, even on these more technical decisions, is important.

Selecting a Random Sample

If the evaluator is sampling to save costs and the ultimate desire is to describe a larger group, probability sampling, in which each unit in the population has a known probability of being selected, may be in order. With simple random sampling—a type of probability sampling—each unit has an equal and independent chance of being selected. Samples drawn in this method, if large enough, are more likely to represent the population than samples drawn through convenience or purposive sampling. Most large assessment projects (e.g., the U.S. National Assessment of Educational Progress [NAEP]) and public opinion polls use probability sampling.

What does probability sampling involve? First, let us define a few terms. A "sampling unit" is an element or collection of elements in the target population. Sampling units could be individuals; classrooms, offices, departments or like units; or entire institutions such as schools, hospitals, or prisons. Care must be taken to select a sampling unit that is consistent with the element about which one would like to make inferences. That is, if we want to draw conclusions about individual schools, we should use schools as the sampling unit, not classrooms or individual pupils.

A "sampling frame" is the list, map, directory, or other source in which sampling units are defined or listed and from which a set of units is selected. If the target population were all small elementary schools (fewer than 200 children) in Iowa, our sampling frame would be a list or directory of those schools. In selecting a sampling frame, the evaluator should consider the degree to which the sampling frame includes all the population of interest. (Are all elementary schools with fewer than 200 children in Iowa included in the list? Have some new schools opened since the list was developed?) Conversely, it is important to determine whether the sampling frame includes units that are not currently part of the population of interest. (Does the list contain schools that have grown to be larger than 200 pupils since the publication of the document?) The degree to which the sampling frame includes the entire target population, and no others, obviously influences the accuracy of the sampling process—the extent to which the final sample represents the population of interest to the evaluation.

To draw a random sample, evaluators must first define the population of interest for the evaluation and specify the sampling unit. Then, they must find the sampling frame that contains all the population, and no others, in the unit of interest. Some adjustments to the sampling frame may be required to exclude units no longer in the population and to add units that are new to the population. If simple random sampling is to be used, evaluators could then use a table of random numbers to select those in the sampling frame from which information will be collected. More typically, the computer can generate a list of random numbers to avoid the trouble of manually selecting numbers. The final numbers, then, are those individuals or units from which data or information will be collected.

A common variant of simple random sampling used in surveys is stratified random sampling. Stratified random sampling is used when evaluators are interested in examining differences among subgroups in the population, some of which are so small that they may not be represented in sufficient numbers in a simple random sample. (If all subgroups of interest are large, stratifying is unnecessary. Sufficient numbers will be attained through random sampling.) Thus, if evaluators were examining parents' attitudes about schools, they might stratify the sample on the dimension of whether or not parents had children in special-needs classes. Such stratifying would help ensure that such parents were sufficiently represented for the evaluators to describe the attitudes of this important subgroup with confidence that they represent the population of parents with special-needs children in the district. Often, samples are stratified for race or ethnicity if it is believed that racial or ethnic minorities may have different opinions and if one or more groups represent only a small proportion of the population. Company surveys might stratify for level of position to make sure that administrators are sampled in sufficient numbers. Stratified random sampling divides the population into *strata* representing the subgroups of interest. Simple random sampling is then used to select units within each stratum.

A comprehensive discussion of sampling appears in Henry (1990). We also recommend that the evaluator study the well-designed sampling procedures used by large-scale assessment projects such as the NAEP or survey studies such as those conducted by the Institute for Social Research at the University of Michigan. Many existing sampling designs can be adopted or adapted by the evaluator.

Figure 15.2 summarizes some key steps for selecting an appropriate sampling method.

Using Purposive Sampling

As noted in Chapter 14, not all sampling procedures are used to obtain a representative sample of the population. When conducting interviews or case studies, purposive sampling is used to select individuals or organizations that can provide the desired information for the study. The goal is to select people who are either informed (experts) on an issue or who represent a particular group that is important to answering an evaluation question. In the latter case, the goal is better understanding of a particular subgroup. In such a case, evaluators must determine the

FIGURE 15.2 ***Checklist for Selecting a Sampling Procedure***

1. Is the design a case study? (If not, proceed to step 2.)
   a. What unit is of interest (individuals, classrooms or offices, organizations)?
   b. How many units are necessary to achieve the desired depth and understanding?
   c. Should the units represent a range of characteristics, a typical unit, or unusual units on one or more dimensions?
   d. What characteristics are of interest in selection?
   e. By what means will the cases be identified?
2. Is generalizing the results beyond the sample a priority?
   a. How would you define the sampling unit (type of unit, critical characteristics)?
   b. Is there a sampling frame that might contain most, if not all, of the sampling units? How can units in the sampling frame that do not represent the sample be eliminated? How can you identify units that may have been omitted from the sampling frame?
   c. Are you interested in examining the responses of small subgroups? If so, consider stratifying to ensure obtaining a sufficient number of the subgroup.

types of cases that should be explored for the evaluation and find ways to identify and select those cases.

With purposive sampling, a sample is drawn based on particular purposes or judgments. Students who are deemed the greatest discipline problems by teachers might be selected to describe the types of discipline problems that teachers encounter. Or a group of "typical" clients in a budget-counseling program might be selected for in-depth interviews to determine the types of problems they encounter in applying information from the program. Purposive samples might be drawn from individuals, units, or organizations that have experienced great success to learn more about the factors influencing their success. For example, high schools that have low dropout rates for at-risk students might be a useful sample in a needs-assessment study exploring strategies for the future. In the evaluation of a program, identifying students or clients who have succeeded and learning more about their progress and the ways in which the program influenced that success could be quite useful for program revision and improvement.

When using purposive sampling, the first step is to identify the characteristics of the purposive sample and to document the rationale for studying them. What types of students or clients will be studied and why? How will collecting information from this subset of people served by the program help answer the evaluation questions?

Before drawing the sample, evaluators should consider how many they want to study. Purposive samples often require extensive data collection on each individual through interviews with those selected and possibly others familiar with their behavior. Thus, data collection with purposive samples can be costly. If so, the number selected may be small. Evaluators should determine how many they can afford to sample and collect data from, and still draw useful conclusions.

Then, evaluators must consider how the cases will be identified. Typical strategies include provider or teacher nominations. But evaluators can also use analyses of existing data to identify potential candidates, peer recommendations, or even observation if the context lends itself to that. Document the procedures and criteria used for selecting cases so that the reader can understand the context. Involve other stakeholders in considering the cases to select. Which cases will best illustrate and add to knowledge on the issue of concern? How can those cases best be identified?

Note: Purposive sampling is not haphazard. Purposive sampling has a purpose; a subgroup has been identified and a rationale has been developed for studying them. Convenience sampling is haphazard sampling. Data are collected from those who happen to be around. If we were to select the first four people coming into an agency for our sample, regardless of who they were, we would be drawing a convenience sample. There are reasons why these people come at this time and we do not know those reasons. Therefore, the sample is neither representative nor purposive. Evaluators can chat with program participants, parents, or family members when visiting the program and such efforts can provide useful information for further investigation, but they should not be considered effective sampling methods. Instead, such conversations provide hints about what to explore in the future with more care and depth.

Cost Analysis

Most program managers are not econometricians and should not be expected to be skilled in identifying all the financial, human, or time costs associated with programs they operate. That leniency cannot extend to evaluators, however, for their evaluation may require them to bring precise information on costs to the attention of developers, deliverers, and administrators who are responsible for their products or programs.

Analyzing costs and benefits for public-sector programs can be a complex undertaking. Public administrators, elected officials, and the public at large are very concerned with the cost of public programs today. Therefore, cost studies are important. However, it is essential to distinguish among the different types of cost studies that can be conducted. Each type is useful but serves different questions, choices, and program stages. We have found Levin and McEwan's (2001) discussions of cost-benefit, cost-effectiveness, cost-utility, and cost-feasibility analyses to be a useful guide to what is possible.

Cost-Benefit Analysis

Cost-benefit analysis makes use of monetary data to compare alternative interventions or programs. The costs of each program are determined and the benefits are identified and monetized or converted to dollar amounts. With such data, cost-benefit ratios can be devised for each intervention or program and the ratios can be compared. Such activities are helpful in calling attention to the monetary value

of program outcomes. Policymakers, no more than any other consumer, should not simply purchase the cheapest program. Instead, they should select programs that provide the greatest dollar impact or output compared with their costs, if those costs are affordable.

Conducting a cost-benefit study essentially involves identifying all costs and all benefits associated with a program and translating any nonmonetary costs or benefits into dollars. While determining all costs and monetizing them can be difficult, monetizing benefits is often more problematic. The outcomes of most public-sector programs are difficult to convert to dollar terms. What is the monetary value of better mental health? Clean air? An additional year of education? One less murder? Educational benefits are often translated into projected gains in earnings or into the amount of money one would have to pay for educational services if they were not provided. Other outcomes that contribute to greater longevity (health programs, clean air) or greater productivity (training, better mental health) also make use of earnings to monetize benefits. Benefits for national parks have been monetized by determining the amount people pay to travel and visit them (Mills, Massey, & Gregersen, 1980). The evaluator is advised to review the literature on cost-benefit studies in the discipline of the program to be evaluated to identify the commonly accepted benefits used and the means for converting these benefits to dollars.

The disadvantage in cost-benefit analysis, of course, is that it can be very difficult to translate all benefits into dollar terms. While gains in earnings are one benefit of education, other benefits are accrued through the impact of education on the quality of life and the educational aspirations of the next generation, to name only two. Further, cost-benefit studies can involve quite technical issues, using discounting to put all costs in the same timeframe ($1,000 in 1970 is not worth the same amount in 2010) and opportunity costs to convey the costs of not pursuing other options. (Yes, you may earn more after going to college, but you must also consider the income lost due to not working full-time during that period of attending school, and lesser seniority and experience once one enters the job market.) These methods can improve the accuracy of the final ratio but add further to the complexity and estimation or judgment involved in conducting such a study. Levin and McEwan (2001) caution that cost-benefit analysis should be used only "when the preponderance of benefits could be readily converted into pecuniary values or when those that cannot be converted tend to be unimportant or can be shown to be similar among the alternatives that are being considered" (p. 15).

One of the most important factors for evaluators to convey to stakeholders about cost-benefit studies is that, in spite of the fact that the study ends with nice, neat numerical ratios, these numbers are quite fallible. Many judgments and estimates are involved in determining the costs and benefits to be included and how to transform these costs and benefits into dollars. Good studies often present several ratios (called sensitivity analysis) to show how changes in assumptions will change the ratios.

But cost-benefit analyses can make powerful statements about the value of programs. Levin and McEwan (2001; Levin, 2005) provide several examples of successes. The cost-benefit study comparing adults who had participated in the

Perry Preschool Program years before with those from a control group who had not participated in the program provided strong evidence for the effectiveness of the program. The cost-benefit study showed benefits in schooling, higher earnings, reduced public assistance, and less involvement in the criminal justice system that monetized to close to $100,000 per participant. For society, the benefit-cost ratio showed a return of almost $7 for each dollar spent on the preschool program (Barnett, 1996). Powerful evidence for the economic returns of preschool programs! Similarly, a study by Levin concerning the use of dietary supplements to reduce anemia in developing countries found a return of from $4 to $38 for each dollar spent in providing supplements. Benefits were generally measured by greater work output because of less anemia (Levin, 1986). The benefit-cost ratios for different methods of providing supplements helped these poor countries (Indonesia, Kenya, and Mexico) determine which strategy would be most beneficial for them to use.

The term "cost-benefit" has become popular and, on more than one occasion, the authors have been asked to conduct a cost-benefit study when, in fact, such a study would not address the information needs of the client. Often, simple cost analyses will suffice to satisfy the client. Given their costs, cost-benefit studies are only cost-effective when stakeholders are trying to make summative decisions about programs with quite different outcomes. Should we rebuild the playground or purchase new books? Which program deserves more funding: public television or children's immunizations? When a choice is to be made among programs with like outcomes, other types of cost studies that do not require monetizing benefits can be more appropriate.

Cost-Effectiveness Studies

Cost-effectiveness analysis involves comparing the costs of programs designed to achieve the same or similar outcomes. When the task for the administrator or stakeholder is to choose from among several different ways to achieve the same goal, this method is the correct choice. Like cost-benefit analysis, cost-effectiveness analysis results in a ratio. However, the benefits side of the ratio is not expressed in monetary terms. Instead, it is expressed as one unit of outcome that would be desired for the programs being compared. The outcome might be one additional year of life, one year's increase in reading ability, an employment or college placement, or one less violent crime. The ratio then shows the cost of each program per outcome achieved. Programs can then easily be compared for their cost-effectiveness in achieving the desired outcome.

The advantage of cost-effectiveness analysis is that benefits do not have to be converted to monetary terms. In addition, cost-effectiveness ratios more appropriately reflect the decisions that most administrators have to make, namely, which program to pursue to achieve a particular goal. However, compared to cost-benefit analysis, there are disadvantages. Only programs with common goals and common measures of effectiveness can be compared using this method. The final ratio does not inform us as to how the program costs are offset by the benefits. In other words,

with a cost-benefit ratio, we are provided with information on whether the benefits are more than the costs. We cannot make this assessment with cost-effectiveness ratios. Finally, cost-benefit studies allow us to convey many, if not all, of the benefits in one ratio. Each cost-effectiveness ratio conveys the costs for only one benefit.

Because many programs have multiple goals, judgment is involved in determining the goal to focus on in the cost-effectiveness ratio. Several ratios may be calculated to reflect the different goals of the programs. A cost-effectiveness study of two reading programs might, quite appropriately, calculate one ratio with the outcome of gains in reading ability and another ratio with the outcome of books read voluntarily in the next year to measure the success of the programs in instilling a desire to read. While several ratios may complicate the decision to be made, they can be useful in conveying the comparative values of programs. An advantage of cost-benefit ratios is that they have the potential to include all program benefits or outcomes on the benefit side of the ratio through monetization. The cost-effectiveness study must develop different ratios for each benefit. However, if the benefits are difficult to translate to monetary terms and the program has only two or three major outcomes, several cost-effectiveness ratios may be preferable.

We hope this brief discussion of cost-analysis provides a sufficient overview to help the reader understand basic approaches and necessary steps. Extensive discussions of cost-analysis of education may be found in Levin and McEwan (2001) and Kee (2004). Yates (1996) discusses ways to conduct cost-effectiveness and cost-benefit analyses in human service settings and provides useful examples of its application to substance abuse, suicide prevention programs, residential programs, and other settings. Williams and Giardina (1993) provide an interesting discussion of cost-benefit analysis as approached internationally; they include examples from the areas of health and transportation. Layard and Glaister (1994) review methods and problems in cost studies using cases in the environmental, health, and transportation fields. Scott and Sechrest (1992) discuss cost studies from Chen's theory-driven approach.

Major Concepts and Theories

1. Evaluators use many designs and methods in their evaluation studies. The choice should be based on the methods that are most appropriate for answering the question at hand, the context of the program, and the values and information needs of the stakeholders. Often, that requires multiple methods.

2. Descriptive designs are the most common designs in evaluation and serve many useful purposes. Cross-sectional designs provide useful quantitative information on large numbers of individuals and groups. Case studies are invaluable for exploring issues in depth, providing thick descriptions of programs in implementation, different outcomes, contextual issues, and needs and perspectives of various stakeholders. Time-series designs are effective in describing changes over time.

3. If the purposes of an evaluation question are causal, evaluators should carefully discuss expectations and implications with stakeholders. Design choices include experimental, quasi-experimental, and explanatory case study designs.

4. Designs can and often are mixed to serve the purposes of the evaluation and the needs of stakeholders.

5. To answer most evaluation questions, data will be gathered from the entire population because the population of interest is relatively small, and external validity—or generalizability—beyond the group of interest is not a priority. Methods of random sampling can be used when the group is large and generalizability is important. Purposive sampling is useful when the evaluation can benefit from information from a particular, identified subgroup to learn more about that group. Purposive sampling is often used in conducting intensive interviews and in case studies.

6. Cost studies help determine if the program outcomes are worth the cost. The most common methods are cost-benefit analysis, used to compare programs with different outcomes; and cost-effectiveness analysis, which can be useful in comparing the costs of programs with like outcomes.

Discussion Questions

1. What designs are most commonly used in your field? What are the strengths and weaknesses of these designs? Having read about many different types of designs, which do you think could be used more frequently in your organization?
2. Some people argue that random assignment is unethical, that everyone should receive the benefits of a new program. What are the arguments against this position? In what circumstances would you feel comfortable with random assignment? Not comfortable? Why?
3. What types of designs do you think should be used most frequently? Why?

Application Exercises

1. Check your evaluation plan [work sheets] from Chapter 14. Are there plans that you would like to change, using what you just learned in this chapter? Would you want to reconsider design or sampling issues? Add cost-related questions? Approach things a little differently?
2. Consider a problem or issue in your organization that is currently controversial. Which design or mix of designs would be most appropriate to address that issue? What evaluation question(s) would your design answer? How would you implement the design?
3. Find a cost-benefit study in your field. Read it and consider the assumptions made. How were benefits quantified? What costs were considered? Whose perspective was used in the ratio (the client, the public)? Were sensitivity analyses conducted? What types of decisions was the study to serve? Would cost-effectiveness analysis have been a more appropriate approach?

4. Find an evaluation study that uses one or more of the evaluation designs or sampling strategies reviewed in this chapter. How does the method or methods shed light on the program? What types of questions do the methods answer? (Example: Read "Of snakes and circles: Making sense of classroom group processes through a case study," a classic case study by Valerie Janesick published in *Curriculum Inquiry,* 12, pp. 161–185, in 1982.)
5. Read Chapter 2, "Crafting Mixed-Method Evaluation Designs" by Caracelli & Greene, in Greene, J. C., & Caracelli, V. J. (Eds.), *Advances in Mixed-Method Evaluations* (1997). (See citation under "Suggested Readings.") Use Caracelli and Greene's framework to critique your own evaluation plan as developed in Question 1 and the study you reviewed in Question 4. What is the primary purpose of your evaluation? Of the one you critiqued? How would you modify your design, or the one you read, based on Caracelli and Greene's discussion and examples?

Relevant Evaluation Standards

We consider the following evaluation standards, which are listed in their entirety in Appendix A, to be relevant to this chapter's content:

U3–Negotiated Purposes
U4–Explicit Values
U6–Meaningful Processes and Products
F2–Practical Procedures
F3–Contextual Viability
F4–Resource Use
P1–Responsive and Inclusive Orientation
P3–Human Rights and Respect
A2–Valid Information
A3–Reliable Information
A4–Explicit Program and Context Descriptions
A6–Sound Designs and Analyses
E1–Evaluation Documentation

Case Studies

For this chapter, we recommend reading one or two of the following five interviews that illustrate different aspects of design and sampling: *Evaluation in Action,* Chapters 2 (Riccio) or 4 (Bickman), 3 (Greene), 5 (Fetterman), or 6 (Rog).

In Chapters 2 and 4, Riccio and Bickman discuss how they each applied a true experiment with random assignment to welfare reform and mental health treatments, respectively, to establish causality in high-profile situations with significant importance. Although the primary focus of their studies is quantitative, they make use of descriptive designs to describe the programs they are evaluating. The journal sources for these chapters are: Fitzpatrick, J. L., & Riccio, J. (1997). A dialogue about an award-winning evaluation of GAIN: A welfare-to-work program. *Evaluation Practice, 18,* 241–252. Also, Fitzpatrick, J. L., & Bickman, L. (2002). Evaluation of the Ft. Bragg and Stark County systems of care for children and adolescents: A dialogue with Len Bickman. *American Journal of Evaluation, 23,* 67–80.

In Chapters 3 and 5, Greene and Fetterman describe their evaluations, which are concerned with program effectiveness, but make use primarily of case study approaches, with the case being the program, and use extensive observation to learn more about the program in practice and to explain its effects. Greene uses a purposive sampling strategy to select individuals who are most likely to be able to implement the conflict-resolution strategies that they learned in the

training in their jobs and then conducts intensive case studies of these trainees' projects. Through these, she learns much about the application of the training and her discussion indicates how these cases informed the evaluation. The journal sources for these chapters are Fitzpatrick, J. L., & Greene, J. (2001). Evaluation of the Natural Resources Leadership Program: A dialogue with Jennifer Greene. *American Journal of Evaluation, 22*, 81–96. Also, Fitzpatrick, J. L., & Fetterman, D. (2000). The evaluation of the Stanford Teacher Education Program (STEP): A dialogue with David Fetterman. *American Journal of Evaluation, 20*, 240–259.

In Chapter 6, Rog describes a national, multisite evaluation study of homeless families. Her purpose is initially causal, to learn how the intervention programs at different sites impacted homeless families. Although her purpose and methods change, she is able to reach some causal conclusions with the type of information she collects. Her design might be viewed as having a combination of descriptive and causal elements. The journal source is Fitzpatrick, J. L., & Rog, D. J. (1999). The evaluation of the Homeless Families Program. A dialogue with Debra Rog. *American Journal of Evaluation, 20*, 562–575.

Suggested Readings

Greene, J. C., & Caracelli, V. J. (Eds.). (1997). Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms. *New Directions for Program Evaluation*, No. 74. San Francisco: Jossey-Bass.

Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.

Julnes, G., & Rog, D. J. (Eds.) (2007). Informing federal policies on evaluation methodology: Building the evidence base for method choice in government sponsored evaluation. In *New Directions for Evaluation*, No. 113. San Francisco: Jossey-Bass.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications*. Thousand Oaks, CA: Sage.

O'Sullivan, E., Rassel, G. R., & Berner, M. (2003). *Research methods for public administrators* (3rd ed.). White Plains, NY: Longman.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Stake, R. E. (1995). *The art of case study research.* Thousand Oaks, CA: Sage.

Yin, R. K. (2009). *Case study research: Design and methods.* (4th ed.). Thousand Oaks, CA: Sage.

16

Collecting Evaluative Information: Data Sources and Methods, Analysis, and Interpretation

Orienting Questions

1. When and why do evaluators use mixed methods for data collection?
2. What criteria do evaluators use to select methods?
3. What are common methods for collecting data? How might each method be used?
4. How can stakeholders be involved in selecting and designing measures? In analyzing and interpreting data?
5. How does analysis differ from interpretation? Why is interpretation so important?

In the previous chapters, we described how evaluators work with stakeholders to make important decisions about what evaluation questions will serve as the focus for the study and possible designs and sampling strategies that can be used to answer the questions. In this chapter, we discuss the next choices involved in data collection: selecting sources of information and methods for collecting it; planning procedures for gathering the data; and, finally, collecting, analyzing, and interpreting the results.

Just as with design and sampling, there are many important choices to be made. The selection of methods is influenced by the nature of the questions to be answered, the perspectives of the evaluator and stakeholders, the characteristics of the setting, budget and personnel available for the evaluation, and the state of

the art in data-collection methods. Nevertheless, using mixed methods continues to be helpful in obtaining a full picture of the issues. Remember, an evaluator's methodological tool kit is much larger than that of traditional, single-discipline researchers, because evaluators are working in a variety of natural settings, answering many different questions, and working and communicating with stakeholders who hold many different perspectives. As in Chapter 15, we will discuss critical issues and choices to be made at each stage and will reference more detailed treatments of each method.

Before discussing specific methods, we will again comment briefly on the choice between qualitative and quantitative methods, in this case referring specifically to methods of collecting data. Few, if any, evaluation studies would be complete if they relied solely on either qualitative or quantitative measures. Evaluators should select the method that is most appropriate for answering the evaluation question at hand given the context of the program and its stakeholders. They should first consider the best source for the information and then the most appropriate method or methods for collecting information from that source. The goal is to identify the method that will produce the highest quality information for that particular program and evaluation question, be most informative and credible to the key stakeholders, involve the least bias and intrusion, and be both feasible and cost-effective to use. Quite simple, right? Of course the difficulty can be in determining which of those criteria are most important in answering a specific evaluation question. For some, the quality of evidence might be the most critical issue. For others, feasibility and cost may become critical.

Qualitative methods such as content analysis of existing sources, in-depth interviews, focus groups, and direct observations, as well as more quantitative instruments such as surveys, tests, and telephone interviews, should all be considered. Each of these, and other methods that are more difficult to categorize, provide opportunities for answering evaluative questions. In practice, many methods are difficult to classify as qualitative or quantitative. Some interviews and observations are quite structured and are analyzed using quantitative statistical methods. Some surveys are very unstructured and are analyzed using themes that make the data-gathering device more qualitative in orientation. Our focus will not be on the paradigm or label attached to the method, but rather on how and when each method might be used and the nature of the information it generates.

Common Sources and Methods for Collecting Information

Existing Documents and Records

The evaluator's first consideration for sources and methods of data collection should be existing information, or documents and records. We recommend considering existing information first for three reasons: (1) using existing information can be considerably more cost-effective than original data collection; (2) such information is nonreactive, meaning it is not changed by the act of collecting or analyzing it,

whereas other methods of collecting information typically affect the respondent and may bias the response; (3) way too much information is already collected and not used sufficiently. In our excitement to evaluate a program, we often neglect to look for existing information that might answer some of the evaluation questions.

Lincoln and Guba (1985) made a useful distinction between two categories of existing data: documents and records. Documents include personal or agency records that were not prepared specifically for evaluation purposes or to be used by others in a systematic way. Documents would include minutes or notes from meetings, comments on students or patients in their files, organizational newsletters or messages, correspondence, annual reports, proposals, and so forth. Although most documents consist of text or words, documents can include videos, recordings, or photographs. Because of their more informal or irregular nature, documents may be useful in revealing the perspectives of various individuals or groups. Content analyses of minutes from meetings, newsletters, manuals of state educational standards, lesson plans, or individual notes or correspondence can help portray a true picture of events or views of those events. One of the advantages of documents is that they permit evaluators to capture events, or representations of those events, before the evaluation began, so they are often viewed as more reliable than personal recall and more credible to outside audiences (Hurworth, 2005). Text documents can be scanned onto the computer and analyzed with existing qualitative software using content analysis procedures. (See "Analysis of Qualitative Data" at the end of this chapter.)

Records are official documents or data prepared for use by others and, as such, are typically collected and organized more carefully than documents. Many records are computerized. Some are collected and held by the agency primarily for internal use, but they are more official than documents and, therefore, are collected more systematially. Such records could include personnel data on employee absences and turnover or data on patients or students and the services they receive, their demographics, test scores, health status, attendance, and such. Other records are organized by external agencies to be used for tracking and in research by others. These would include test scores collected by state departments of education, measures of air quality collected by environmental agencies, economic records maintained by a variety of government agencies, census data, and the like. Of course, such public data can be useful for giving a big picture of the context, but they are rarely sensitive enough to be used to identify a program effect. Remember that, although existing information can be cheaper, the cost will not be worth the savings if the information is not valid for the purposes of the current evaluation study. Unlike data collected originally for the study, this information has been collected for other purposes. These purposes may or may not match those of your evaluation.

Identifying Sources and Methods for Original Data Collection: A Process

In many cases, existing data may be helpful but cannot serve to completely answer the evaluation questions to the satisfaction of stakeholders. Therefore, for most studies evaluators will have to collect some original data. In Chapter 14, we

reviewed the typical sources of data, that is, the people from whom one might collect information. Recall that common sources for data include:

- Program recipients (e.g., students, patients, clients, or trainees)
- Program deliverers (social workers, therapists, trainers, teachers, physicians, nurse practitioners)
- Persons who have knowledge of the program recipients (parents, spouses, coworkers, supervisors)
- Program administrators
- Persons or groups who might be affected by the program or who could affect its operation (the general public, future participants, organizations or members of interest groups involved in the program)
- Policymakers (boards, CEOs, elected officials and their staffs)
- Persons who planned or funded the program (state department officials, legislators, federal funding agency officials)
- Persons with special expertise in the program's content or methodology (other program specialists, college or university researchers)
- Program events or activities that can be observed directly

To select a source and method, evaluators take these steps:

1. Identify the concept or construct that must be measured in each evaluation question specified in the evaluation plan. For example, if the question is: "Did patients' health improve after a six-week guided meditation class?" the key concept is patients' health.

2. Consider who has knowledge of this concept. Several sources may emerge. Of course, patients are likely to have the most knowledge of how they feel, but with some conditions, they may not be able to accurately report their condition. In these cases, family members or caregivers may be an important secondary audience. Finally, if patients have a specific medical condition, such as high blood pressure, high cholesterol, or diabetes, evaluators may also look to the medical providers or existing records to obtain periodic physical measures. In this case, multiple measures might be very useful to obtain a fuller picture of patients' health.

3. Consider how the information will be obtained. Will evaluators survey or interview patients and their family members or caregivers? How much detail is needed? How comparable do the responses have to be? If evaluators want to compare responses statistically, surveys may be preferable. If they want a better sense for how patients feel and how the program has or has not helped them, interviews may be a better strategy. Both surveys (of many) and interviews (perhaps with a subset) may be used but at a higher cost. Similarly, with the patients' blood pressure or other measures, should evaluators obtain the records from files, or do they also want to talk to the health care providers about the patients' health?

4. Identify the means to be used to collect the information. Will surveys be administered face-to-face with patients as they come to the meditation class or to

another office? Will they be conducted as a short interview by phone? If patients or their family members are likely to use computers, an electronic survey might be used. Finally, surveys could be mailed. Some of these choices have to do with the condition of the patients. Family members or caregivers may be less accessible, so face-to-face administration of survey items or interviews would have to be arranged, and visits to homes might not permit family members to discuss the patient privately. Telephone interviews might introduce similar privacy and validity concerns. Mailed or electronic surveys might be preferable.

5. Determine what training must take place for those collecting the data and how the information will be recorded.

Another concept to be considered in the evaluation question is the six-week guided mediation class. The evaluators need to determine if the program is being delivered as planned and perhaps assess its quality. The program theory proposes that this class should take place in a particular way, under leaders with certain types of training and skills, for a certain length of time, and with specified equipment and facilities. A quiet, carpeted room with comfortable places for sitting might be important. In measuring program delivery, evaluators need to identify the critical concepts to monitor or describe, and then use a similar process to identify data sources, methods, and procedures.

We have briefly described and illustrated a process for considering the construct(s) to be measured, potential sources for the construct, and then, given the source, the best manner for collecting the information. Let us move on to discussing different types of data collection, and their strengths and weaknesses, so readers can make a more informed choice about the method(s) to use.

Using Mixed Methods. Note that, as shown in the previous example, evaluators often use mixed methods. In using mixed methods to measure the same construct, evaluators should consider their purposes in order to select the right mix and order for the measures. Mixed measures might be used for any of the following reasons:

- Triangulation, or increasing the validity of the measurement of the construct by using sources and methods with different biases. If these different measures show the same results, evaluators can be more confident that they have validly captured the construct. Our use of both patients' self-reports of their health and family members' or caregivers' reports is for triangulation, with the expectation that family members' reports should generally confirm those of the patient.

- Complementarity, or measuring different facets of the construct with different measures or sources to increase our understanding of the construct. The comparison of self-reports of health with the medical measures is complementary, designed to shed light on different, though related, aspects of health. For example, blood pressure may have improved in some patients, but they still feel poorly. Others may feel more energetic or relaxed or in some way healthier even though their blood pressure levels have not declined. Each concept—perceptions of health and physical

measures of a health indicator—are important and inform evaluators' views of patients' overall health.

- Development purposes, when responses to one measure help evaluators in developing the next measure. In these examples, interviews and surveys may be used for development purposes. Interviews may first inform the types or wording of survey questions. The analysis of the survey data may then be followed by interviews with patients or health care providers to learn more about trends found in the survey data.

In the next sections, we will review some common methods for collecting information. We will be building on the classification scheme that we introduced in Chapter 14 (see pp. 348–349.) Our focus here, however, is providing more detail on particular methods that you have chosen to use, describing their strengths and weaknesses, and providing information on some other choices evaluators make in implementing particular methods.

Observations

Observations are essential for almost all evaluations. At minimum, such methods would include site visits to observe the program in operation and making use of one's observational skills to note contextual issues in any interactions with stakeholders. Observation can be used more extensively to learn about the program operations and outcomes, participants' reactions and behaviors, interactions and relationships among stakeholders, and other factors vital to answering the evaluation questions. Observation methods for collecting evaluation information may be quantitative or qualitative, structured or unstructured, depending on the approach that best suits the evaluation question to be addressed.

Observations have a major strength: Evaluators are seeing the real thing—the program in delivery—a meeting, children on the playground or students in the halls, participants in their daily lives. If the evaluation questions contain elements that can be observed, evaluators should definitely do so. But, many observations also have a major drawback—the fact that the observation itself may change the thing being observed. So, evaluators may not be seeing the real thing but, instead, the way those present would like to be observed. Program deliverers or participants may, and probably do, behave differently in the presence of a stranger. Some programs or phenomena being observed are so public that the observers' presence is not noted; the program has an audience already. For example, court proceedings or city council hearings can be observed with little or no concern for reactivity because others—nonevaluators—are there to observe as well. But, in many cases, the presence of the evaluator is obvious and, given the circumstances, observers may need to be introduced and their role explained. Those being observed may have to give informed consent for the observation to take place. In such cases, it is recommended that several observations be conducted and/or that observations continue long enough for those being observed to become more accustomed to the observation and, therefore, behave as they might without the presence of

the observer. Evaluators should judge the potential for reactivity in each setting being observed and, if reactivity is a problem, consider how it might be minimized or overcome. We will discuss reactivity more later.

Of course, observations in any evaluation may be confidential or require informed consent. Evaluators should consider their ethical obligation to respect the dignity and privacy of program participants and other stakeholders in any observation.

Unstructured Observations. Unstructured methods are especially useful during the initial phase of the evaluation. Evaluators can use their observational skills to note critical features during their first interactions with stakeholders. How do clients and other stakeholders respond to the evaluator? How do they interact with one another? Who is invited to what types of meetings? Who represents different stakeholder groups? Jorgensen (1989) writes:

> The basic goal of these largely unfocused initial observations is to become increasingly familiar with the insiders' world so as to refine and focus subsequent observation and data collection. It is extremely important that you record these observations as immediately as possible and with the greatest possible detail because never again will you experience the setting as so utterly unfamiliar (p. 82).

Unstructured observations remain useful throughout the evaluation if evaluators are alert to the opportunities. Every meeting is an opportunity to observe stakeholders in action, to note their concerns and needs and their methods of interacting with others. If permitted, informal observations of the program being evaluated should occur frequently.[1] Such observations give the evaluator a vital picture of what others (e.g., participants, deliverers, administrators) are experiencing, as well as the physical environment itself. Each member of the evaluation staff should be required to observe the program at least once. Those most involved should observe the program frequently to note changes and gain a greater understanding of the program as it is delivered. When two or more members of the evaluation team have observed the same classes, sessions, or activities, they should discuss their perspectives on their observations. All observers should keep notes to document their perceptions at the time. These notes can later be arranged into themes as appropriate. (See Fitzpatrick and Fetterman [2000] for a discussion of an evaluation with extensive use of program observations or Fitzpatrick and Greene [2001] for a discussion of different perceptions by observers and how these differences can be used.)

[1]We say "if permitted" because some program interactions may be private, e.g., therapy sessions or physical exams in health care. Such sessions may ultimately be observed but not in an informal fashion without participants' consent. By "informal observation" we mean wandering into the program and observing its general nature. Training or educational programs and some social services and judicial programs present this opportunity.

Structured Observations. Structured methods for observation become useful when the evaluator desires to observe specific behaviors or characteristics. What specific behaviors or characteristics might be observed? For many public-sector programs, critical characteristics may be physical in nature—the size and arrangement of classrooms, park maintenance, road quality, playground facilities, library collections, physical conditions and/or density of program facilities, and so forth.

Other observations can involve interactions between program deliverers and participants—teacher-student interactions, teacher-administrator interactions, student-administrator interactions, physician-nurse-patient interactions, social worker-client interactions, therapist-client interactions, receptionist-client interactions, and so on.

A final category of observations is participants' behaviors. What behaviors might one observe? Imagine a school-based conflict-resolution program designed to reduce playground conflicts. Observations of playground behaviors provide an excellent method for observing outcomes. Imagine a new city recycling program for which there are questions about the level of interest and nature of participation. The frequency of participation and the amount and type of refuse recycled can be easily observed, although the process may be a little messy! Students' attention to task has been a common measure observed in educational research. While many programs focus on outcomes that are difficult to observe, such as self-esteem or preventing alcohol and drug abuse many others lead to outcomes that can be readily observed. This is particularly the case when the target audience or program participants are congregated in the same public area (e.g., hospitals, schools, prisons, parks, or roads).

Structured methods of observation typically involve using some type of form, often called observation schedules, to record observations. Whenever quantitative observation data are needed, we advise reviewing the literature for existing measures and, if necessary, adapting this instrument to the particulars of the program to be evaluated. Other concerns in structured observation involve training observers to ensure consistency across observers. Careful training and measuring for consistency or reliability are critical. The differences that may emerge in the data should be based on real differences in what is observed, not differences in the observers' perceptions. It is important to consider not only what to observe but also the sampling for observation: Which sites should be observed? Which sessions or what times should be observed? If individual participants or students are selected from among a group for observation, how will they be chosen? (See Greiner, 2004, for more on structured observations, particularly training observers, calculating inter-rater reliability, and using it for feedback to improve the quality of the observations.)

Use of Observations in Evaluation. A major use of observations in evaluation is to examine and evaluate the implementation of the program. Many evaluations are concerned with describing the way in which a program is implemented. Evaluations of outcomes or impacts should be preceded by a study of program implementation so that evaluators know what the program is like in implementation. Without such a description, evaluators cannot know the reasons for program success or failure because the program delivery is a "black box." They don't know what was actually

delivered. But evaluations of the program process are also valuable simply as process studies to describe how the program was delivered and to make recommendations for improvement. Such studies can be particularly useful when the program model has been proven to be effective in previous research and, therefore, implementation of the model in specific ways is important. If the model is adapted, it is important for evaluators to examine why the adaptations have occurred, such as to meet the needs of different types of students or participants than encountered in past research (perhaps good reasons!) or because of program deliverers' lack of training, resources, or time (not-good reasons!). Although logs or diaries from program deliverers and self-reports from participants provide some information about program delivery, observation is a critical method for collecting data on program implementation.

Brandon et al. (2008) provide an excellent example of using observation to measure the quality of program implementation. In their case, they were evaluating the implementation of an inquiry-based science program in middle schools. They identify three purposes for examining implementation: adherence of the implementation to the original model, the amount of exposure students or participants have to the model (dosage), and the quality of the implementation. Adherence is whether programs are delivered according to a logic model or program theory; quality is concerned with how the program is implemented. They note that "Observations are necessary for measuring quality because they avoid self-report biases by providing external reviewers' perspectives" (2008, p. 246). They make use of videotapes of "key junctures" in the inquiry-based curricula and focus on teachers' questioning strategies. As they note, the process forces them to make decisions about what to observe and, thus, to "winnow down [a] list of program features to examine the essential characteristics that most directly address program quality and are most likely to affect student learning and understanding" (p. 246). They must make decisions about which schools and teachers to videotape and what features or events to record, and then they must train judges to compare teachers' responses through the videotapes. The project illustrates the complexity of observation, but also its utility in evaluating and identifying what constitutes quality performance. Readers' use of observation may be less complex, but it can be informed by the procedures and forms used by Brandon et al. and their focus on quality.

Observations can also be used as quality checks on adherence in program implementation. In such cases, the factors to describe or evaluate should be key elements in the program's logic models or theory. Program deliverers may be directed to maintain logs or diaries of their activities, and participants may be asked to report what they have experienced in regard to the key elements. But mixed methods, including observation, are useful to document the reliability and validity of these self-report measures. Evaluation staff can train observers to document or describe the key elements of the program using checklists with notes to explain variations in what is observed. Zvoch (2009) provides an example of using classroom observations of teachers delivering two early childhood literacy programs across a large school district. They used observation checklists of key elements in the program and found variation in program implementation both at early and later stages of the program. They were able to identify teacher characteristics and

contextual variables associated with adherence to the program model that were then helpful in analyzing and identifying problems. For example, teachers with larger class sizes were less able to adhere to the model. Not surprising, but helpful to know for future program dissemination.

Surveys

Surveys (sometimes referred to as questionnaires)[2] are used in evaluation for a wide variety of purposes. Braverman (1996), in his review of surveys in evaluation, notes that, "Surveys constitute one of the most important data collection tools available in evaluation" (p. 17). Surveys are used when the number of sources is too large for cost-effective interviewing and there is a need or desire to have information from many individuals and to analyze it in a quantitative manner. Common uses include the following:

- Surveys of program participants upon completion of the program, or during the program, to obtain their views and reports on the program activities. Reports are more factual—respondents are describing what happened to them in the program. Views are more evaluative, with respondents commenting on the quality of the program and their reaction to the utility of various components for them as individuals.[3]
- Surveys of program deliverers (teachers, trainers, health care workers, any employees of the program who interact with participants to deliver services) to learn about how they delivered the program, any adaptations or changes they made and the rationale for those changes, their perceptions of the participants and their needs, participants' reactions and behaviors during the program, changes they observe in participants, and their recommendations for change. Program deliverers have expertise in the program as they delivered it, knowledge of other programs or methods, and knowledge of program participants. When there are many deliverers, surveys may be used to obtain their input, although these are often supplemented with more detailed interviews with a smaller sample of program deliverers.
- Surveys of program participants, or their family members (parents, spouses) to learn about program outcomes. These respondents can sometimes more objectively report participants' behaviors than participants themselves, who have spent time in the program and may be invested in seeing change.
- Surveys of intended target audiences for the program to learn of their perceived needs or their behaviors, knowledge, skills, or attitudes that may be the focus of the program or their characteristics (education, employment, age, location, family status, etc.).

[2]Survey more appropriately refers to the general method, whereas questionnaire, interview protocol, and the like refer to instruments used to collect the actual data.

[3]Most organizations make use of satisfaction surveys with clients, parents, and the like. Often these are conducted in a rote and superficial way. We encourage evaluators and administrators to make use of these surveys to add items that might be useful for a particular, timely issue.

- Surveys of stakeholders or the general public to obtain perceptions of the program or of their community and its needs or to involve the public further in policy issues and decisions regarding their community (Henry, 1996).

These are some of the common uses of surveys, but they can be used for many purposes with many different audiences. We will now move to how evaluators identify or develop surveys for evaluation studies.

Identifying an Existing Survey. As with any type of information to be collected, the evaluator should first consider whether there are existing questionnaires that would be appropriate to use in the current study. The review of literature conducted in the early stages of the study should have helped the evaluator to identify common surveys or survey items used to assess the construct of interest. Although surveys are rarely included in journal articles, authors are typically willing to share their items if contacted. Evaluation reports typically do include, in an appendix, surveys that were used. Finally, there are references that contain measures that may be useful. *The Mental Measurements Yearbook* series has published independent reviews of commonly used tests and other measures for many years. Today it is available online at http://www.unl.edu/buros/bimm/html/00testscomplete.html. An introduction to the references may be found at www.unl.edu/buros/.

Developing Surveys. When the purpose of the survey is to measure opinions, behaviors, attitudes, or life circumstances quite specific to the program being evaluated, the evaluators are likely to be faced with developing their own questionnaires. In this case, we recommend developing a design plan for the questionnaire that is analogous to the evaluation design used for the entire evaluation. In the first column, list the questions (not the item) to be answered by the survey. That is, what questions should the results of this survey answer? In the second column, indicate the item type(s) that should be used to obtain this information. A third column may be used after items are developed to reference the numbers of the items that are linked to each question. A fourth column can then specify the means of analysis. Table 16.1 provides an illustration. This design then becomes a

TABLE 16.1 ***Sample Design Plan for Questionnaire***

Question	*Item Type*	*Item Number*	*Analysis*
1. What are clients' opinions of the agency?	Likert 5-point scale	2–20	Descriptive for each item and total score
2. How did clients first learn of the agency?	Multiple-choice	21	Percentages
3. What type(s) of services do they receive from the agency?	Checklist	22–23	Percentages
4. Do opinions differ by type of service required?		Score on 2–20 with 22–23	t-tests and ANOVA, explore

guide for planning the questionnaire and analyzing the information obtained. It helps evaluators confirm that they have included a sufficient number of items to answer each question. (Some questions require more items than others.) The design also helps to avoid items that sound interesting but, in fact, don't really address any of the evaluation questions. Evaluators may decide to include such items, but their purpose should be further explored. Items that do not answer a question of interest lengthen the questionnaire and show disrespect for the time and privacy of the respondents.

In selecting a type of item, consider that many variables can be measured with several different item formats. Most questionnaires solicit relatively structured responses and are typically analyzed statistically. Item types include multiple-choice items; items with adjectival responses (e.g., rating items using a five-point scale of excellent [1] to poor [5]); items with adverb responses (always, frequently, etc.); Likert-scale items (agree-disagree) and open-ended items. Open-ended items may be short answer, such as number of children who live at home, or may require longer responses that can be summarized qualitatively by themes, for example, best elements of a program or recommended changes to be made. Surveys with many open-ended items will suffer from poor response rates or absence of responses on these items unless the respondents are very interested in the subject. Today, electronic surveys and the ease of keyboarding or typing can increase responses to such open-ended items, but they can still present a problem. The following is a general list of constructs that might be measured with a survey and the type of item evaluators might use:

- Attitudes: Likert-scale items
- Behaviors: Adverbs (frequency of behaviors) or multiple choice (types of behaviors)
- Opinions: Adjective items (ratings of favorability) or multiple choice (selecting preferences)
- Life status or circumstances: Multiple choice (with numeric ranges, yes-no, or alternatives)

Careful development of the questionnaire draft, instructions, and cover letter (if distributed by mail) then follow. In developing the first draft, evaluators should consider these issues:

1. Sequencing questions
   - **a.** Are later responses biased by earlier questions?
   - **b.** Does the questionnaire begin with easy, nonthreatening, but pertinent questions?
   - **c.** Are leading questions (ones that lead to a certain response) avoided?
   - **d.** Is there a logical, efficient sequencing of questions (e.g., from general to specific questions; use of filter questions when appropriate)?
   - **e.** Are closed- or open-ended questions appropriate? If closed, are the categories exhaustive and mutually exclusive? Do responses result in the desired scale of data for analysis (i.e., nominal, ordinal, interval)?

   f. Are the major issues covered thoroughly but minor issues passed over quickly?
   g. Are questions with similar content grouped logically?
2. Wording questions
   a. Are questions stated precisely? (Who, what, when, where, why, how? Don't be too wordy.)
   b. Does the questionnaire avoid assuming too much knowledge on the part of the respondent?
   c. Does each item focus on only one issue?
   d. Is the respondent in a position to answer the question, or must he make guesses? If so, are you interested in his guesses?
   e. Are definitions clear?
   f. Are emotionally tinged words avoided?
   g. Is the vocabulary at the reading level of the audience? If any technical terms, jargon, or slang are used, are they the most appropriate way to communicate with this audience?
   h. Are the methods for responding appropriate? Clear? Consistent?
   i. Are the questions appropriately brief and uncomplicated?
3. Establishing and keeping rapport and eliciting cooperation
   a. Is the questionnaire easy to answer? (Questions are not overly long or cumbersome.)
   b. Is the time required to respond reasonable?
   c. Does the instrument look attractive (i.e., layout, quality of paper, etc.)?
   d. Is there a "respondent orientation"?
   e. Does the cover letter provide an explanation of purpose, sponsorship, method of respondent selection, anonymity?
   f. Is appropriate incentive provided for the respondent's cooperation?
4. Giving instructions
   a. Are the respondents clearly told how to record their responses?
   b. Are instructions for returning the survey clear? If the survey is sent by mail, is a stamped return envelope provided?

Many evaluators make use of the Total Design Method developed by Donald Dillman to develop surveys. His most recent book includes information on using the Internet to conduct surveys, as well as traditional methods (Dillman, Smyth, & Christian, 2009). He provides specific suggestions for formatting, length, cover letters, follow-ups, and inducements to increase response rates. Other useful references on developing and conducting surveys include Fink's fourth edition of *How to Conduct Surveys* (2008) and Cox's book (2007), which focuses on developing surveys in education.

Surveys can be administered face-to-face, by telephone, electronically, or in paper-and-pencil format delivered through the mail or distributed in person. The previous comments are concerned mainly with paper-and-pencil questionnaires, which can be mailed or, ideally, administered to participants at a time when they are together in a group. Next, we will cover collecting survey-type data by telephone, electronically, or face-to-face.

Telephone Surveys or Interviews. As many of you know—especially those who have been greeted by a phone call from a professional survey firm during the middle of cooking dinner and juggling children and homework—telephone interviews have become a much more commonly used method of collecting information. While the information obtained from telephone interviews can be qualitative and analyzed in such a manner, in practice a telephone interview is typically more akin to a questionnaire administered orally than to a personal interview. Like a paper questionnaire, it should be brief to encourage participation. Unlike the personal interviewer, the telephone interviewer has difficulty in establishing rapport due to the lack of eye contact and other nonverbal cues, although there is certainly more potential for rapport, encouragement, and responding to questions than with a paper or electronic survey. While branching is often used in telephone interviews to skip questions that are inappropriate for the current respondent, the interviewer is seldom encouraged to adapt questions as in an unstructured personal interview. Instead, standardization is encouraged.

The primary differences between administering a telephone survey and a paper questionnaire are these:

1. The visual element is removed in a telephone survey; that is, respondents cannot read the items. This presents some advantages because respondents cannot read ahead or skip around, nor can they change their responses for consistency. The order is prescribed by the telephone interviewer.

2. The oral element is introduced in a telephone survey. Questions are asked orally and people respond orally. Therefore, items that are long or complex, such as multiple-choice items or rating items with many options, must be broken down into smaller parts when asked by a telephone interviewer. However, open-ended items can be more frequent in telephone interviews than on paper questionnaires because respondents are more willing to speak a sentence or even a paragraph than to write one.

3. Information can be obtained more quickly using telephone interviews than questionnaires distributed through the mail, but if questionnaires can be distributed to intact groups (employees, students, clients), questionnaires may be the more efficient and effective method.

4. If items are probing for sensitive information, paper surveys may be preferable because respondents feel more anonymous. Although telephone interviews are commonly conducted using random-digit dialing, those answering may believe their telephone number is recorded or feel the need to impress the interviewer. For example, Dillman and Tarnai (1991) found that people were 11% more likely to report driving under the influence of alcohol when responding to a paper survey than when asked by telephone.

5. Telephone interviews can make use of branching (skipping items that are not relevant to a particular respondent). If a paper survey looks quite long, but many items will be skipped by respondents, a telephone interview using branching might

be preferable. Telephone interviewers typically make use of a computer program that automatically moves to the branch of items appropriate for the respondent given his or her previous responses. For example, a telephone survey of parents with school-age children might branch to different items for parents responding about a child in kindergarten or preschool than for parents with a child in high school.

6. The costs of each method can be relatively similar if paper surveys are mailed. The categories of costs are simply different. Costs for mailing questionnaires include clerical time, paper, copying, postage, postcards, and envelopes. Costs in telephone surveys are primarily the costs of staff involved in conducting and monitoring the surveys. Long-distance charges, rental of facilities for telephoning and training, and the purchase of phones may also be factors. Many evaluators and researchers make use of companies that conduct telephone surveys and already have trained interviewers. This is much more efficient than recruiting and training interviewers on one's own. Telephone survey companies also have software for conducting interviews and can make recommendations about wording items and provide instructions in ways that are most appropriate for telephone interviews.

Telephone surveys should be considered preferable to mailed surveys when (a) there is a need for speed, (b) respondents may be reluctant or unable to complete paper surveys but can be reached by telephone, and (c) the questions lend themselves to being answered over the telephone.

Electronic Surveys. Electronic surveys have become a common means for administering surveys. Today, so many people in industrialized countries have personal access to the Web and e-mail that there is much less concern than in the past that this method of distribution may bias the sample of respondents. Nevertheless, evaluators should always consider whether the group they hope to survey will have access to e-mail or a web site that links them to a survey and check their e-mail or the web site relatively frequently. In surveying company employees or college students, evaluators can be sure that group members have been provided with an e-mail account and are expected to use it to receive information. Other more diverse groups may not have such access. If a subgroup does not have access, evaluators can choose a mixed approach, using e-mail with some and mailed surveys with others, or they can distribute all surveys by mail. Some key issues include the following:

1. Electronic surveys are more like paper surveys than like telephone interviews. Again, respondents are able to see the items, so complex items can be used, but, typically, respondents are not permitted to skip ahead or move back and, therefore, cannot change items to be consistent. Respondents can usually type faster than they can write, so interested or motivated respondents may reply more frequently and more extensively to open-ended items on electronic surveys than on paper surveys, but less than on telephone interviews where their response is oral.

2. Studies have shown that response rates to electronic surveys are somewhat lower than to mailed surveys (Fricker & Schonlau, 2002). Many now advocate using a mixed-mode approach. For example, Converse, Wolfe, Huang, and Osward (2008) found that using a mailed announcement of a web-based survey led to a higher response rate than an e-mail with a link to the web-based survey.

Evaluators typically find that directing respondents to a web site that contains the survey makes the respondents, particularly if they are employees or clients, feel more anonymous, since their e-mail addresses are not linked to their responses. Several companies provide web sites that guide questionnaire development and provide basic analyses of responses. These include surveymonkey.com, zoomerang.com, questionpro.com, and hostedsurvey.com.

Face-to-Face Surveys. Surveys are administered by an interviewer for a variety of reasons: to gather information from clients who have literacy problems or may have difficulty understanding the questions, to stimulate or motivate participants to respond, or to permit occasional probing by the interviewer to increase the quality and nature of the response. Conducting surveys face-to-face is more costly than self-administered surveys and can be more costly than telephone surveys if visits must be made to the respondents' homes. Nevertheless, this method of data collection remains a viable option if respondents are coming to the agency or organization or to a meeting where the survey can be administered. Of course, those administering the survey should ensure that a private place, preferably a separate room, is available so that respondents' anonymity, privacy, and dignity are respected. Face-to-face surveys are more akin to telephone interviews than to paper or electronic surveys because voice is being used. Some of the primary differences between face-to-face surveys and phone interviews are:

1. Interviewers can potentially establish greater rapport with respondents in person than by telephone. To do so, interviewers need to be well trained in methods for making respondents comfortable and establishing rapport without biasing the nature of the responses. So, while the interviewer should be making eye contact, explaining purposes of the survey, answering questions, smiling, making the respondent feel comfortable, and encouraging responses, the interviewer should not be revealing his or her personal preferences or showing approval or disapproval for certain types of responses.

2. Consideration should be given to how to record responses in a natural, unobtrusive, but open way. The interviewer might quickly mark selected responses, but if the responses are anticipated to be more open-ended, tape-recording the interview should be considered. (Evaluators differ in their views regarding the use of tape recorders. Some feel it frees the interviewer to make more eye contact and establish rapport, while still documenting the dialogue; others prefer hand-writing notes, believing that the presence of the tape recorder can inhibit discussion by intimidating the interviewee. We think tape-recording is more useful to accurately capture

the response, to maintain eye contact and rapport, and to allow the interviewer to focus on the questions and the respondent rather than on writing responses.)

3. Respondents are more likely to give socially desirable responses in face-to-face interviews than in any other format. This tendency can be reduced somewhat by matching respondents and interviewers on gender, age, race/ethnicity, and other potentially important characteristics, because respondents are more likely to gain rapport and give more honest and complete responses to interviewers they perceive to be like themselves. A well-trained interviewer, appropriately matched to the respondent, can be more successful at eliciting more detailed responses from some stakeholder groups than a self-administered or telephone survey might.

4. If interviews are to be conducted by several people, careful consideration should be given to training. Interviewer bias can be a major source of error (Braverman, 1996). Sources of error can include inconsistent use of probes, body language, verbal comments from the interviewer, or changes in the wording of questions. So, interviewers need to be trained in standardized methods for delivering the questions including using probes, pauses, and prompts; methods for recording the information; and means for establishing and maintaining rapport (Bernard, 2000).

Interviews

Interviews are a central part of qualitative data collection. Observations are typically the core element of qualitative evaluation, but there is much that the evaluator cannot observe and, even when observing, the evaluator's perspective differs from that of others experiencing that same phenomenon. Therefore, qualitative interviews are used for learning the perspectives, attitudes, behaviors, and experiences of others. Stake notes that "The interview is the main road to multiple realities" (1995, p. 64). In other words, only through hearing and interpreting the stories of others through interviews can the evaluator learn the multiple realities and perspectives that different groups and individuals bring to an object or experience.

A major difference between surveys and collecting data through personal interviews is that interviews allow clarification and probing and permit exploration and discovery. Interviews are useful when the nature of the information to be collected is more ambiguous and greater depth is needed than would be permitted with the more structured items of a survey. Personal interviews require more time than surveys, so they can cost more if many people are to be interviewed. However, they can provide a wealth of information.

Good interviewing, however, is a skill. While good interviewers encourage people to talk and tell their stories, they also guide the discussion, through questions and probes, to learn more about the evaluation questions of interest. At the same time, as McCracken notes, interviewers must demonstrate that they are "a benign, accepting, curious (but not inquisitive) individual who is prepared and eager to listen to virtually any testimony with interest" (1988, p. 38). Kvale recommends that evaluators consider their role and purpose and provides two metaphors of the interviewer: the interviewer as a miner, who uses interviews to "unearth" knowledge that

may be facts or "nuggets of essential meaning" (p. 3), and the interviewer as a traveler who "wanders through the landscape and enters into conversations with the people encountered" (1996, p. 4). The traveler-interviewer reflects the more qualitative approach. The metaphor reveals the role of interviewers in wandering, learning, possibly changing, and returning to interpret what they have learned to people in their home country. Kvale develops his concept of interviews as conversations designed to learn and understand the "life world" of the person being interviewed in respect to the object being evaluated. He discusses and provides numerous examples of interviews to achieve this purpose.

Rubin and Rubin (2004) discuss interviews as conversational partnerships, recognizing that interviews are a social relationship in which norms and power relationships must be given attention. They discuss how to develop the partnership of an interview, designing main questions and probes and preparing follow-up questions.

Although the questions asked in each qualitative interview will differ, Stake (1995) reminds us that:

> The interviewer needs to have a strong advance plan. It is terribly easy to fail to get the right questions asked, awfully difficult to steer some of the most informative interviewees on to your choice of issues. They have their own. Most people are pleased to be listened to. Getting acquiescence to interviews is perhaps the easiest task in case study research. Getting a good interview is not so easy (p. 64).

When planning your questions, consider the evaluation question you are trying to answer. What information do you need to answer the question? What experiences or opinions do you want them to describe? What linkages, thoughts, or experiences of theirs do you want to elicit? To explore and probe? Develop a brief list of broad questions and be prepared with prompts.

One of the most common errors of an interviewer is to talk too much. After rapport is successfully established, the interviewer is present primarily to listen and encourage responses. A good interviewer should become comfortable with pauses and not feel compelled to fill gaps hurriedly. Respondents often pause to convey difficult or sensitive information. If the interviewer rushes in to break the silence, such information is lost. Similarly, the interviewer should be prepared with prompts and phrases to continue discussion: "tell me more about that," "that's interesting," "oh, yes," or, even "uh-huh" show you are listening and encourages the interviewee to continue without the interviewer determining the direction. Reflecting on the last statement of the interviewee can be helpful in encouraging the respondent to continue. Interviewers should be careful not to add their own interpretation, though, but seek only to learn more about the meaning given by the person being interviewed.

Following are some helpful hints for developing interview questions:

1. Begin with simple, informational or "chatty" questions to establish rapport and learn more about the style and manner of the interviewee.

2. Keep the language pitched to the level of the respondent. Questions posed to specialists can rely on the terminology with which they are familiar and show

the interviewer's own expertise on the issue. This use of technical jargon can encourage the specialist to talk in more depth, but questions posed to the general public must use language more commonly understood. Special care should be taken in interviewing individuals or groups whose language may differ from the typical evaluator's frame of reference. Agar (2000) discusses the language/culture he discovered within an organization in an ethnographic evaluation of a tuberculosis screening program.

3. Avoid long questions. They often become ambiguous and confusing.

4. Consider whether you are seeking facts, opinions, or broader perspectives with each question and phrase your question accordingly. Use prompts or follow-up questions to obtain the type of information you are seeking.

5. Do not assume that your respondent possesses factual or firsthand information. Parents may be able to report what books their children read, but only the children can tell you accurately how much they enjoy reading.

6. Listen for implicit assumptions or biases that may be as important as the answers to your questions. Consider whether the interviewees' comments suggest certain orientations or perspectives and decide whether to probe those perspectives. For example, if the interviewee is complaining about school vouchers, the interviewer might want to probe to see whether the interviewee is against the vouchers for personal or for political reasons and, once that is determined, to learn more about the nature of these reasons.

7. Decide whether you need a direct question, an indirect question, or a combination. An example of a direct question is, "Do you ever steal on the job?" An indirect question might be, "Do you know of anyone ever stealing on the job?" A combination might be, "Do you know of anyone ever stealing on the job?" followed by "Have you ever taken anything while on the job?"

8. Frame the question so that, to the degree possible, you communicate what you want to know. For example, if interested in reader preferences in magazines, don't ask, "How many magazines do you read?" Ask, "Which magazines do you read?"

9. Protect your respondent's ego. Don't ask, "Do you know the name of the Chief Justice of the Supreme Court?" Ask, "Do you happen to know the name of the Chief Justice of the Supreme Court?"

10. If you are interested in obtaining negative, or critical, information, give your respondent a chance to express her positive feelings first so that she feels comfortable being critical. First ask, "What do you like about X?" Then ask, "What don't you like about X?" or "What bothers you about X?"

A final issue in interviewing concerns how to record the information obtained. As noted previously, in face-to-face surveys evaluators disagree on whether to use tape recorders or to take notes during the interview. For qualitative

interviews, in which eye contact and developing rapport are critical to helping the respondent feel comfortable in telling her story, decisions about procedures should be made very carefully. Taking extensive notes can make respondents uncomfortable and certainly detracts from the body language that would encourage honesty and sharing of experiences. If tape recording is inappropriate because it would also reduce respondents' comfort levels, jot a few notes or key words as the interview proceeds. Then, immediately after the interview, make extensive notes on the respondent's remarks. Share these notes with the respondent to ensure his or her agreement (or disagreement) with your interpretation.

There is disagreement on this issue. Stake (1995) writes:

> Getting the exact words of the respondent is usually not very important, it is what they mean that is important. . . . Interviewees often are dismayed with transcripts not only because of the inelegance of their own sentences but because they did not convey what they intended. And the transcript arrives long after context and innuendo have slipped away (p. 66).

On the other hand, others view taping interviews as the norm and even recommend videotaping to capture nonverbal cues in certain circumstances (Kvale, 1996; McCracken, 1988; Patton, 2001). Patton (2001) writes:

> No matter what style of interviewing you use and no matter how carefully you word questions, it all comes to naught if you fail to capture the actual words of the person being interviewed. The raw data of interviews are the actual quotations spoken by interviewees. Nothing can substitute for these data: the actual things said by real people (p. 380).

Taping permits the interviewers to study the remarks, provides more detail for data analysis, and remains to attest to the veracity of the process.

Denzin and Lincoln (2005), Kvale (1996), Patton (2001), and Rubin and Rubin (2004) all provide useful guidance to conducting qualitative interviews. These sources are highly recommended for evaluators who plan on using interviews to collect information.

Focus Groups

Focus groups have become an increasingly popular method of obtaining qualitative information from a group of individuals. Focus groups are like an interview in that they involve face-to-face interaction, but they build on the group process. A skilled focus group facilitator will make use of ideas or issues raised by participants in the focus group to obtain reactions from others in the group. Discussion in focus groups is not always interviewer to interviewee, but often among participants themselves. Thus, the interview is very much a group process.

Focus group techniques emerged from the field of marketing, where they were used to gauge potential customers' reactions to new products and to learn

more about customers' needs and wants in regard to the product. Focus group methods have now been adapted to many different settings to obtain information on how individuals react to either planned or existing services, policies, or procedures or to learn more about the needs and circumstances of participants or potential clients. In addition to reacting to issues, focus group participants may suggest new methods or describe circumstances that pose problems with existing programs or policies.

Focus groups are particularly useful in needs assessments and monitoring studies and for formative evaluations. Participants can describe their experiences with or their reactions to proposed new programs or changes, the changes they would recommend, and any beliefs, attitudes, or life circumstances they have that might facilitate or hinder the success of the program. Focus groups can help confirm or disconfirm program theories during the planning stages of programs. They can raise novel ideas based on participants' own experiences. Focus groups can also be useful in discovering more about program outcomes, such as how participants have used what they gained, what barriers they faced, or what changes they would make in the program.

Focus groups typically consist of six to ten individuals who are relatively homogeneous, but unknown to each other (Krueger & Casey, 2009). These small numbers help stimulate group interaction and provide them all with a means to express themselves and their concerns. The goal is not representativeness but obtaining in-depth information. Responses from individuals in the focus group may prompt others in the group to reveal more. Homogeneity is desirable to facilitate group interaction; noticeable differences in education, income, prestige, authority, or other characteristics can result in hostility or withdrawal by those who are lower on those dimensions. Typically, three or four focus groups may be conducted on the same subject matter to determine if new issues arise, but combining 20 people into one focus group will not serve that purpose!

The role of the leader is to facilitate discussion by introducing and describing the process, posing initial and periodic questions, moderating the responses of more vocal members, encouraging responses of quieter members, and monitoring the time to ensure that critical questions are covered. The leader may also ask questions to clarify ambiguities or obtain reactions from other group members. The questions should, of course, lead to answers that inform the evaluation questions for the study, but they should be questions that encourage the participants to converse and provide opinions and specific examples of experiences (Krueger, 2005).

Fontana and Frey (2000) note that the skills required for leading a focus group are similar to those required for a good interviewer, but the leader must also be knowledgeable about methods for managing group dynamics. Leading an effective focus group can be a challenging task. One frequent error in focus groups is to rely too extensively on short, forced-choice questions (e.g., yes or no) or to have group members respond by raising hands. The focus group then really becomes a structured group interview, not a focus group, because it has lost the

key focus group characteristics of member interaction, openness, and exploration. A successful moderator can also help a participant feel comfortable voicing a minority view and, thus, avoid "group think."

In selecting a moderator, consider how that moderator's characteristics and background can enhance or impede group discussion. Employees, or someone known to the focus group participants, should never be used as a focus group leader. The position or attitudes of the leader can influence such discussions in undesirable ways. It can be desirable, though not always necessary, to match moderator and group characteristics on critical demographic variables such as age, gender, race, or ethnicity. At minimum, the moderator should have a good knowledge of the culture or lifestyles of the participants in order to understand and interpret comments and effectively facilitate interactions.

Groups are typically led by one moderator with an assistant to observe body language and interactions and to assist in interpreting the session. Sessions are usually tape-recorded, and participants are reimbursed for their time. Sessions generally last one-and-a-half to two hours. The environment for the focus group is important. Generally, refreshments are available and the room is arranged to be conducive to conversation. Results are interpreted through analysis of transcripts from tapes. The results may be analyzed by themes in more open-ended discussions or by responses to groups of questions posed by the moderator.

For more information on focus groups, see Krueger and Casey (2009) or Barbour (2008).

Tests and Other Methods for Assessing Knowledge and Skill

Tests are a common method for collecting evaluative information in education and training programs. Knowledge acquisition is often the primary objective of educational programs, and the acquisition of knowledge is generally, but not always, measured by tests. Evaluators in other fields also make use of tests, though less extensively than educational evaluators. Evaluators in training settings may use tests, though their ultimate objective is often application on the job or impact on the organization. Those in the health field may use tests for the many educational programs conducted for clients or for health education programs for practitioners. Evaluators in social services might use tests to measure outcomes in employment or parenting programs. Therefore, all evaluators need to have some knowledge of tests as a method of data collection.

The most common approaches to testing are norm-referenced testing (NRT), criterion-referenced testing (CRT), objectives-referenced testing (ORT), and domain-referenced testing (DMT). These four strategies have many elements in common, but depending on which strategy is chosen, the procedures for test development and interpretation can be quite different. Norm-referenced tests are intended principally to compare students' performances against others taking the same test. They are administered in many school districts to assess progress, although their use has declined with the mandate for states' standards-based tests.

The California Achievement Test, the Comprehensive Test of Basic Skills, and the Iowa Test of Basic Skills are common examples of norm-based tests. The strength of such tests is that they permit comparison with established norm groups. They can be helpful in answering questions such as, "How is our school doing in conveying commonly accepted knowledge and skills compared with other schools in the nation?" Their chief weakness for evaluation purposes is that the content may not be valid for the curriculum or program being evaluated.

In contrast to norm-referenced tests, criterion-referenced tests are developed specifically to measure performance against some absolute criterion. The standards tests used by each of the 50 states in the United States are criterion-referenced to the standards each state has developed. They have typically replaced norms-based tests for assessment purposes because citizens and policymakers are more concerned with assessing what students know rather than how they compare with others. (See Chapter 2 for a discussion of "A Focus on Measuring Outcomes.") Standards-based tests are used by most states to judge the performance of schools or school districts. (See Fitzpatrick and Henry [2000] for a discussion of some of the issues concerning standards-based testing.) As this book goes to press, the U.S. Congress is considering legislation to revise the No Child Left Behind Act that mandated state standards and tests, but it is anticipated that this portion of the Act will continue. Meanwhile, many states continue to reform their standards and their tests to make them more appropriate and valid. In some states, standards have been set so high that only a small proportion of students have had courses covering the test content. Standards are often much higher than the learning achieved by previous generations of students, particularly in math and science.[4]

Scores on these tests may provide useful information for the evaluator working in K–12 settings, but should be judged by the evaluator like any other existing data. Consider whether the items on the test adequately measure the appropriate concepts to answer the evaluation questions at hand. In some cases, the standards-based test can serve as an important and useful measure. Consider also using subscores or even individual items if these portions of the test are better indicators of the constructs stated in the evaluation question. Always remember and remind others that the goal is not to "raise test scores," as we so often hear, but to improve student learning. Identify the knowledge or skills that are the focus of the program or curriculum being evaluated. If scores, subscores, or items on standards-based tests can provide information about that learning, by all means use them! However, we, along with the American Evaluation Association, encourage multiple measures of learning outcomes. Tests can serve as one effective means of assessing learning, but others are needed.

[4]A State Department of Education-sponsored study of the 10th grade math test for the Colorado standards test, the Colorado Student Assessment Program (CSAP), found that the content of the 10th grade math test was more comparable to sophomore college math than 10th grade math curricula. The cut point for "proficient" corresponded to a point above the 90th percentile on a nationally normed test. No wonder very few 10th grade students in Colorado scored at the "proficient" level!

Traditionally, criterion-referenced tests were designed to address a particular program or curriculum. Scores on the test can be used to judge students' progress on the curriculum and, ultimately, as one method for judging the success of the program. These measures are useful for many program evaluations because the content of the items reflects the curricula.

Objectives-referenced tests make use of items keyed to specific instructional objectives in an educational or training program. Such tests are useful for formative evaluation feedback for teachers or trainers to help them examine areas in which objectives are being achieved and areas that need improvement. They can also be useful for summative decisions about program success in achieving certain learning goals. Domain-referenced tests are used to estimate a student's knowledge of a specific content or domain. The items are not linked to a curriculum but rather to the content domain being measured (e.g., American history, comparative anatomy). These items, too, can be useful for evaluative purposes, though such tests are costly to develop compared with objectives-referenced and criterion-referenced tests. Typically, evaluators would select previously developed and validated measures. Domain-referenced tests can be used to answer questions such as "How much do our graduates know about X content?" Standards can be developed to reflect the school's or organization's expectations regarding the amount of knowledge a graduate or a student finishing a course should have.

In addition to paper-and-pencil tests, other methods exist to measure knowledge, as illustrated in the list of data-collection methods in Chapter 14. These include performances of skills through simulations or role-playing or samples of work contained in student portfolios or work products of employees. As always, evaluators should choose methods that are appropriate to measure the constructs specified in the evaluation question and that are appropriate for the program context and stakeholders' standards of credible evidence. Although standards-based tests have been the most common measures of student learning in the United States in recent years, portfolios continue to be used in many school settings. A portfolio is "a collection of a student's work"; as Mabry notes, "Because of the opportunity to include many examples of work over time, portfolios as an assessment technique offer an unparalleled means of enhancing the validity of inferences of student achievement and of displaying student growth" (2005, p. 323). Although portfolios are used often in assessment, they are less commonly used in evaluation. However, evaluators should consider their use as part of a multiple-measures strategy, as they balance scores on standards-based tests quite nicely because of different biases. They represent real student work products, although they are, of course, less standardized than tests. For evaluative purposes in some skills-based content areas and in training, performance tests can serve as useful evaluative measures. For example, oral speaking skills can be judged or rated by observers. In measuring conversational ability in a foreign language, a structured language proficiency interview would clearly be more appropriate than a paper-and-pencil test. In measuring the skill to use scientific equipment to perform an experiment, a performance assessment in the lab would probably be

most appropriate. In measuring the ability to recognize and correct grammatical and spelling errors, a paper-and-pencil test may be most efficient. The focus on standards-based tests has, unfortunately, decreased the use of these approaches to measuring learning in evaluation in recent years.

As mentioned earlier, *The Mental Measurements Yearbook* may be a useful source for other measures. This series has published independent reviews of commonly used tests and other measures for many years. See the comprehensive online review at http://www.unl.edu/buros/bimm/html/00testscomplete.html. An introduction to other references at the site, including *Tests in Print,* may be found at www.unl.edu/buros/.

Table 16.2 summarizes the data-collection methods we have reviewed in this section and some of the important characteristics of each. Table 16.3 then illustrates methods that might be used to address particular evaluation questions.

TABLE 16.2 ***A Review of Various Means of Data Collection***

Data Collection Method	*Characteristics*
Documents	Nonofficial papers: minutes, notes, plans Reveals actions, thinking, perceptions uninfluenced by the study
Records	Official documents: census, attendance, salaries More valid and reliable than documents
Observation	Observations of program context and activities, participant behaviors, and environments Can be structured or unstructured Useful in some way in almost every evaluation
Surveys	Reports of attitudes, opinions, behavior, life circumstances Can be administered in person or by mail
Telephone Interviews	Purposes are similar to those of a survey; questions can be more open-ended, but must be shorter Can develop rapport and use verbal prompts
Electronic Interviews or Surveys	Questions delivered and answered using computer technology Items may be constructed as open or closed
Interviews	Qualitative interviews are useful for eliciting values, perspectives, experiences, and more detailed responses Can be structured (face-to-face surveys) or qualitative
Focus Groups	A group discussion with 6-10 persons to learn reactions, experiences Useful when group interaction can encourage and enhance responses
Tests	Used to examine knowledge and skills Primarily used in education and training

TABLE 16.3 ***Connecting Data Collection to Evaluation Questions: Some Examples***

Evaluation Question	*Data Source and Method*
What are the primary reasons students miss school?	School records on absences, teacher interviews, counselor interviews, interviews with students with excessive absences and their parents, survey to students
How do sixth-grade teachers implement the conflict-resolution program in their classrooms?	Teachers logs, students surveys and reports, office records, selected observation
What do citizens think of the city-mandated water restrictions? To what extent are citizens obeying the restrictions?	Telephone survey and city water records
Why do good teachers leave the X school district?	Interviews followed by mailed or electronic surveys to larger number (first use archival data to identify good teachers who left)
Did the X reading program succeed in improving students' reading comprehension?	Paper criterion-referenced test and oral test to sample

Planning and Organizing the Collection of Information

Data-collection methods must be sanctioned by the proper authorities. These authorities can include the client, administrators of the program to be evaluated, program staff and participants, Institutional Review Boards (IRBs) or other review committees within the organization. In addition to seeking approval through proper channels and following organizational policies, the evaluator should seek the input of those who will be involved actively or passively in the collection of information (e.g., responding to surveys, helping to administer tests, and observing activities or being observed). These audiences' cooperation can be vital to successful data collection. If they object to the data-collection methods or procedures or fail to understand the purpose, they can sabotage the data-collection by providing false or misleading information or encouraging others to do so. Others simply may not take the data collection seriously. Explaining the importance of their cooperation can prevent many potential problems. For respondents guaranteeing confidentiality or anonymity can be critical. Rewards, such as release time or feedback from the study, may also encourage full cooperation. Adherence to ethical practices that protect participants' rights is also essential to ensure access to data sources.

Technical Problems in Data Collection

The evaluator's version of Murphy's Law goes something like this: "If anything can go wrong in collecting information, it will." A comprehensive list of potential problems would fill this chapter, but here are a few of the major ones:

- Unclear directions lead to inappropriate responses, or the instrument is insensitive or off-target. (Always pilot-test your methods.)
- Inexperienced data collectors reduce the quality of the information being collected. (Always include extensive training and trial runs. Eliminate potential problem staff before they hit the field. Monitor and document data collection procedures.)
- Partial or complete loss of information occurs. (Duplicate save, and backup files and records; keep records and raw data under lock and key at all times.)
- Information is recorded incorrectly. (Always check data collection in progress. Cross-checks of recorded information are frequently necessary.)
- Outright fraud occurs. (Always have more than one person supplying data. Compare information, looking for the "hard to believe.")
- Procedures break down. (Keep logistics simple. Supervise while minimizing control for responsible evaluation staff. Keep copies of irreplaceable instruments, raw data, records, and the like.)

Analysis of Data and Interpretation of Findings

Data Analysis

Evaluations involve processing mountains of information that, if not organized in a form that permits meaningful interpretation, is often worthless or, worse, misleading. The aim of data analysis is to reduce and synthesize information—to make sense of it—and to allow inferences about populations. When considering alternative methods for data analysis, the evaluator should ask two questions:

1. What methods of data analysis are appropriate for the questions I am trying to answer, the information that I plan to collect, and the method I will use to collect information?
2. What methods of data analysis are most likely to be understood and to be credible to the audiences who will receive reports?

Involve stakeholders in data analysis from the beginning. Meeting with the client and other important stakeholders to review results can demystify the data analysis stage and actively involve potential users. The evaluator can learn the types of information that are of most interest to the client or different stakeholders and the most effective ways of presenting that information. Working with the client or group, the evaluator will often learn new questions and issues that the data analysis can address.

The analysis of qualitative and quantitative data today includes many methods. We will alert readers to significant events or methods and recommend

references for further investigation, but, as ours is not a textbook on data analysis, it would be inappropriate for us to imply that we can summarize the vast fields of analysis in a few pages.

Quantitative Data Analysis. For quantitative data analysis, evaluators should consider each evaluation question that is addressed with quantitative data and consider how to summarize those results for each important stakeholder group. If the target audience is researchers or policy analysts and the evaluation question concerns relationships among a number of variables, multivariate methods and other advanced statistics may be used. However, typically, stakeholder groups want quantitative data to be summarized and analyzed in a relatively parsimonious way. Henry (1997), a quantitatively-oriented evaluator, has provided a useful discussion, guide, and examples to using graphs to convey different types of evaluation data.

Analyzing differences in outcomes among particular groups of students or clients and between different classrooms, schools, or sites can be a useful way to learn more about how and why the program works. Pawson and Tilley (1997) recommend doing subgroup analyses to learn more about program theory. For example, learning that the program works at one site and not at another can lead the evaluator to explore differences between the sites that may have contributed to the different outcomes. These discoveries help shed light on what parts of the program are most critical and with whom. This is the type of integrated, iterative evaluation design that leads to a fuller, more complete understanding of the program and its effects. (See, for example, our discussion of Zvoch [2009] under "Observation" in this chapter. He does exactly that.) We also recommend statistics texts by Babbie et al. (2010), which provide guides to using the Statistical Package for the Social Sciences (SPSS), or Salkind (2010), who provides guidance for using Excel to analyze data.

Recent trends in analyzing quantitative data include conveying not only statistical significance, but also effect size. It is the evaluators' responsibility to help clients and other stakeholders understand that statistical significance, while useful, simply conveys the probability that there is some relationship between variables. A smaller value *P* does not mean a greater relationship, but simply more certainty that there is a relationship. As stakeholders and policymakers become more sophisticated users, they have become interested in effect size to provide some indication of the actual magnitude of the program's effect on the outcomes of interest. See Kline (2009) for a practical discussion of effect size, its calculation, and use.

Analysis of Qualitative Data. Stake (1995) observes that qualitative and quantitative techniques are most different from each other at the stage of data analysis. "The qualitative researcher concentrates on the instance, trying to pull it apart and put it back together again more meaningfully—analysis and synthesis in direct interpretation. The quantitative researcher seeks a collection of instances, expecting that, from the aggregate, issue-relevant meanings will emerge" (Stake, 1995, p. 75). As qualitative data are collected, analysis is also beginning. The evaluator is formulating categories, revising categories, reviewing field notes, and collecting more information until different perspectives begin to be more fully revealed. But how do

evaluators summarize the vast amounts of information they have collected? How do they verify it and ensure that it has credibility to the stakeholders and other users? At some point, evaluators begin to consider how to organize and summarize their qualitative data and use it to tell the multiple stories that have emerged.

The method of analysis of qualitative data depends on the nature of the data and the conceptual framework employed in the analysis. Methods for qualitative data analysis concern looking for patterns or themes, developing working hypotheses from these, and then conducting confirmation checks or analyzing negative cases or cases that do not support or strengthen the working hypotheses. Data analysis is iterative, conducted as a process to identify and explore alternative themes. Authors who provide useful details and examples for analyzing qualitative data include Bernard and Ryan (2010), Patton (2001), Strauss and Corbin (1998), and Yin (2009).

Today, several software packages are available for analyzing qualitative data, including NVivo for analysis of text or content (see http://www.qsrinternational.com/); AccuLine, designed specifically for use in schools (see http://www.harpe.ca/Download.php); AnSWR, developed by the Centers for Disease Control for text analysis (see http://www.cdc.gov/hiv/software/answr.htm); and many others. The American Evaluation Association web site provides a list and brief description of software for analyzing qualitative data at http://www.eval.org/Resources/QDA.htm.

Interpreting Data

Data analysis focuses on organizing and reducing the information collected into themes or statistical descriptions and inferences. Interpretation, in contrast, attaches meaning to organized information and uses it to answer the evaluation questions. Analysis may be thought of as organizing and summarizing information; interpretation as applying values, perspectives, and conceptual ability to formulate supportable conclusions. Interpretation is the valuing component of evaluation—the third branch of Alkin's and Christie's evaluation tree, with methods and use as the other two major branches or issues of importance to evaluation. (See Chapter 4 for more on Alkin and Christie's tree.)

Interpretation should be characterized by careful, fair, open methods of inquiry. Anyone who claims that the "numbers speak for themselves" is either naïve or a shyster. Interpretation means using the data to answer the evaluation questions. This includes judging the object of the evaluation and considering the implications of those judgments. Schwandt (2004), who believes that evaluators have placed too much emphasis on methodology, argues that the critical issue in evaluation is the quality of the evaluation judgment. Stake and Schwandt, writing on quality, note:

> Making judgments of quality constitutes a core professional responsibility of evaluators (just as making judgments of health and illness constitute a core imperative of doctors, and making judgments of guilt or innocence is a core imperative of judges). Evaluators shoulder the responsibility of judging quality—a responsibility of getting it right (or of at least making a significant contribution to discussions of getting it right) in the face of the lived reality of the ambiguity of quality and the absence of hard and fast, steady and universal standards for quality (2006, p. 416).

Each evaluator looks at the evaluation information and its analysis, twists it around, discovers nuances, and generates insights—things that others may never have seen without the evaluator's help—in an individual way that affects the outcomes of the evaluation. If evaluation is to serve an educational function, results must be interpreted so that audiences know how best to use or consider them. Stakeholders can, and should, be involved in those interpretations or judgments to facilitate understanding and use, but the ultimate role of evaluators is interpreting the results and reaching a judgment. Stakeholders would not have hired the evaluator unless they wanted the expertise he or she can bring in either reaching a judgment or assisting them in doing so.

Guidelines for Interpreting Findings. Of course, data analyses and the interpretation of findings should be linked to answering the evaluation questions posed in the evaluation plan. These questions represent the information needs of stakeholders and should be answered, to the extent possible, with the evaluative findings. The entire evaluation has been planned around these questions. In some cases, as we noted in Chapter 13, evaluators and stakeholders will have specified criteria and standards for judging the element of the program specified in the evaluation question. In such cases, the specified standards can serve as a guide. How do the data or the information collected help us to answer the evaluation question, to judge the program or program component against the criterion and specified standards? Occasionally, interpretations can be relatively straightforward. The criterion for judging a high school dropout prevention program may have been the proportion of participants who completed their high school degree or remained in school. The standard specified may have been 70%. The data obtained can be compared with that standard and judgments made. If the outcome of the dropout prevention program to be evaluated is 50%, certainly the program has not attained the standard. Is the program a failure? Perhaps. Evaluators need to consider and discuss the evidence to explain this particular implementation of the program with these specific students. Is there some promise that the program might achieve a 70% rate in the future? Is retaining half of the students, who were likely to fail to complete school without the program, a sufficient outcome? How does this outcome compare with other programs of similar intensity and cost? One can see that even with a relatively clear criterion and standard, final interpretations can be difficult.

Other factors enter into interpretation and may be considered given the context of the program and the proclivities of the various stakeholder groups. These include:

1. Determining whether objectives have been achieved
2. Determining whether laws, democratic ideals, regulations, or ethical principles have been adhered to
3. Determining whether assessed needs have been reduced

4. Establishing the value of program or participant accomplishments through cost methods, opinions of policymakers or other stakeholders, or other means
5. Comparing results with those reported by similar entities or endeavors.

But how does one move from the data analyses to these conclusions? Often our mixed methods result in conflicting findings. Such ambiguity should not be viewed as troubling. Instead, it suggests the need for evaluators to take care in reaching their final interpretations. Evaluators might explore the data further, seeking confirmation and consistency from other sources. When evidence is contradictory, consensus should not be forced. Alternative interpretations and honest discussions of the ambiguities can be presented; recommendations can be made for further, focused study on the issue.

Stakeholder Involvement. Input from others is essential to interpretation. Interpretation of data analyses is not the sole province of the evaluator. Most evaluators have learned that interpreting and summarizing results in isolation is generally an unsound practice. The evaluator brings only one of many pertinent perspectives to bear and, in fact, is sometimes less prepared to offer insightful interpretations than others who can look at the data through fresh eyes. Stakeholders closely connected to the program have valuable knowledge and experience and can provide perspective. Clients and participants who have experienced the program can provide understanding.

In fact, stakeholder involvement in interpreting the results of the data analysis can serve several purposes. In addition to potentially adding to the validity or comprehensiveness of the conclusions, their involvement can increase their later use of the information as they understand more about why the conclusions were reached. Finally, stakeholder involvement in interpreting results can build evaluative capacity in the organization. The evaluator's role in facilitating the interpretation can help stakeholders learn that multiple perspectives can and do exist and can help them find ways for sorting out and using those perspectives.

Stakeholders can be called together in meetings of either similar stakeholders or diverse groups. If many participants are involved, evaluators may have them break into smaller groups to discuss their interpretations and reach conclusions. Results may be sent to participants prior to the meeting to save time in presentation and to permit them to reflect on the analyses. Or participants may be given relatively raw data, such as percentages on responses to each survey item or streams of comments on open-ended questions or interviews with confidentiality protected. Participants may then be asked to take a role in analyzing the results in response to each evaluation question. These processes can add to the transparency of the analysis and interpretation process and provide stakeholders with useful skills. But, just as evaluators have biases, so do stakeholders. In acting as facilitators, evaluators must help stakeholders to be data-oriented in reaching their conclusions. (See interview with Jean King in *Evaluation in Action,* where she describes her work with teachers, parents, and educational administrators in facilitating their interpretation of data on the quality of the school district's special

education program. As a facilitator, she requires stakeholders to point to evidence for each of their conclusions making use of the data they have received.)

Finally, the Joint Committee *Standards* (2010) provide guidance for interpretation:

> **A 7** *Explicit Evaluation Reasoning:* Evaluation reasoning leading from information and analyses to findings, interpretations, conclusions, and judgments should be clearly and completely documented.

AEA's Guiding Principle A3 also has relevance:

> Evaluators should communicate their methods and approaches accurately and in sufficient detail to allow others to understand, interpret, and critique their work. . . . Evaluators should discuss in a contextually appropriate way those values, assumptions, theories, methods, results and analyses that significantly affect the interpretation of the evaluative findings.

Thus, through data collection, analysis, and interpretation, evaluators move to answering the evaluation questions posed in the initial plan. The next step, then, is considering how to report the final information, although, as the reader will see in the next chapter, in order to maximize use, reporting is an ongoing process.

Major Concepts and Theories

1. Evaluators make use of many different data sources and methods. The selection of sources and methods is dependent on the nature of the evaluation question(s), the context of the program to be evaluated, and the nature of credible evidence for stakeholders and clients.

2. The evaluator should consider a wide array of methods to collect information. These include documents and records, observations, questionnaires, interviews, focus groups, and tests.

3. Multiple methods can be used to increase validity, to enhance understanding, or to inform subsequent data collection.

4. Quantitative data are analyzed using descriptive and inferential statistical methods as necessary to answer the evaluation question of interest. Qualitative data are analyzed for patterns and themes. Categories are formed and revised as information is accumulated and new considerations emerge.

5. Data must be interpreted to answer the evaluation questions and to provide evidence for the final judgments to be made in the evaluation. Interpretations can be based on bringing together different data sources, methods, and analyses to answer each evaluation question. Clients and other stakeholders can be actively involved in this interpretation to permit different perspectives to emerge, to increase the validity of the interpretation, and to enhance use.

Discussion Questions

1. Should stakeholders or clients be involved in the technical considerations of data collection, analysis, and interpretation? Why or why not? What do they add? What are the strengths and risks of involving them?
2. In studying the implementation of a math program in a high school, would you rather use observations, teacher logs, or student reports to learn more about how the program was delivered? What do you see as the strengths and weaknesses of each method in this instance? If you used mixed methods, which methods would you use?
3. Discuss the advantages and disadvantages associated with collecting observational data. Consider a program you know. What information might be usefully collected through observation? What important program outcomes would be difficult to measure using observational methods?

Application Exercises

1. Examine your worksheets from Chapter 14. How would you revise the sources and methods for data collection that you considered then? What method(s) are most appropriate for answering each evaluation question?
2. Plan an interview to be conducted with fellow students on their reactions to this course. How does it meet their needs? How will they use the information in the future? Design the interview to answer these questions and two others of your own formulation. Develop the questions to ask in the interview. Then interview, individually, three other students. What differences do you find in responses? Does your interviewing style change? Improve? How? Is an interview the best way to answer this question? How does it compare to the use of a survey or focus group? Under what circumstances would you use each approach? Both approaches?
3. Consider your place of work. What documents and records exist there that might be useful for evaluation? How are the documents different from the records? What problems might you encounter in accessing either type of information?
4. In small groups, plan and develop a survey to measure attitudes toward your university. First, develop the questions the survey is to answer. (This may be done as a large group before going into small groups or as part of the small group exercise.) Then, consider the appropriate types of items to answer each question. Develop draft items complete with an introduction and instructions. Pilot-test your questionnaire on another group and discuss their responses and interpretations. How would you revise your instrument? Was a questionnaire the best way to gain this information? Why or why not?
5. What methods of data collection would you use to answer the following questions:
   a. Do the methods used by teachers at Smith High School correspond to the principles of block scheduling introduced last fall?
   b. Did the new reading curriculum result in improved reading abilities among second-graders? Increased interest in reading?

c. What types of recruitment strategies are most effective in involving fathers in Early Head Start?
d. What do different stakeholder groups—parents, students, teachers, counselors, and coaches—think of changing the high school schedule to 9:00 AM to 4:00 PM?
e. What strategies were most useful in helping children of new immigrants in their transition to Haley Middle School?

Relevant Evaluation Standards

We consider the following evaluation standards, which are listed in their entirety in Appendix A, to be relevant to this chapter's content:

U2—Attention to Stakeholders
U4—Explicit Values
U6—Meaningful Processes and Products
U8—Concern for Consequences and Influence
F2—Practical Procedures
F3—Contextual Viability
F4—Resource Use
P1—Responsive and Inclusive Orientation
P3—Human Rights and Respect
A1—Justified Conclusions and Decisions
A2—Valid Information
A3—Reliable Information
A4—Explicit Program and Context Descriptions
A5—Information Management
A6—Sound Designs and Analyses
A7—Explicit Evaluation Reasoning
A8—Communication and Reporting

Case Studies

For this chapter, which covers many different issues concerning data collection and analysis, we recommend several different interviews that may be pertinent to readers' particular interests: *Evaluation in Action,* Chapters 3 (Henry), 5 (Fetterman), 8 (King), 11 (Conner), and 13 (Wallis and Dukay).

In Chapter 3, Henry discusses how he and his advisory group choose many different indicators of school quality to develop a school report card. They make use of existing data, but he discusses ways in which they make the choices to present valid information. He also surveys citizens of Georgia to learn what they would like to know about their schools. Finally, he discusses the visual format he develops to permit parents and citizens to readily interpret the data.

In Chapter 5, Fetterman discusses his use of intensive classroom observation to judge the quality of teaching in the Stanford Teacher Education Program. He and his research assistants become participant-observers, attending every class and sharing information through photographs and Internet discussions.

In Chapter 8, King describes her work with stakeholders to develop, analyze, and interpret surveys used to evaluate the special education program in a school district. King acts primarily as a facilitator to allow stakeholders to gain skills in evaluation and understanding multiple interpretations.

In Chapter 11, Conner discusses using stakeholders to develop a survey and his positive views about that process. He also discusses the role he and his research assistants took in using observation and informal interviews to describe the decision-making processes used in 29 different communities regarding community health.

In Chapter 13, Wallis and Dukay discuss their use of existing psychosocial measures, physical measures, test scores, existing data, interviews, and focus groups to evaluate an orphanage in Tanzania. The interview sheds light on choices concerning cultural competence in regard to selecting measures, training interviewers, conducting interviews, interpreting data, and many other issues that arise when conducting evaluation in a different culture.

Suggested Readings

Denzin, N. K., & Lincoln, Y. S. (2008). *Strategies of qualitative inquiry* (3rd ed.). Thousand Oaks, CA: Sage.

Dillman, D. A., Smyth, J., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Krueger, R. A., & Casey, M. A. (2009). *Focus groups: A practical guide for applied research* (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2001). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Stake, R. E. (1995). *The art of case study research.* Thousand Oaks, CA: Sage.

17

Reporting Evaluation Results: Maximizing Use and Understanding

Orienting Questions

1. What considerations are important in tailoring the reporting of evaluation results to audience needs?
2. What are some ways that results can be communicated to stakeholders?
3. How can a written evaluation report be designed to be most effective for users?
4. How should oral reports of evaluation results be organized and presented?
5. What are some other ways that evaluation results can be disseminated?

In the prior two chapters, we have discussed the collection, analysis, and interpretation of evaluation information. Of course, these activities are not ends in themselves but terribly important means to making evaluation information useful. It seems obvious that such information is not likely to be used effectively unless it has been communicated effectively. Yet reporting is too often the step to which many evaluators give the least thought.

In the past decade, evaluators have realized that it isn't enough to craft a good evaluation report. Indeed, evaluators have become increasingly aware that one can work hard to maximize the quality of one's report and still find that the impact and influence it has on its stakeholders, programs, or policies is at best negligible and, at worst, zero. Thoughtful evaluators now contemplate at the outset how their results might be used and consider ways to ensure that they are useful.

In this chapter, we will review various methods for communicating results to different audiences. These include considering the ongoing nature of reporting and its purposes, recognizing and choosing from the many different methods of reporting, and planning the means of reporting to address the needs and characteristics of particular audiences. We conclude with a discussion of the final written report and oral presentations.

Purposes of Evaluation Reporting and Reports

Evaluators often tend to consider evaluation reporting as the final step. Although we place this chapter at the end of Part Four on Practical Guidelines, evaluators should begin to report on the evaluation well before the end of the project. Traditionally, evaluators have written a final evaluation report, and that has been the focus of reporting—the end of the process. Today, we realize that reporting is about understanding and learning. To maximize the use of evaluation results, we need to talk with primary stakeholders along the way, telling them what we are learning in the evaluation, getting their reactions, and learning what is surprising and what was expected. Reporting entails engaging in meaningful dialogue with the intended primary users so the final results do not come as a surprise. Instead, learning—their learning and ours—has occurred along the way. Results that are surprising tend not to be used. They contradict the intended users' experience and, thus, are forgotten or dismissed as methodologically inadequate (Weiss & Bucuvalas, 1980b). By engaging in an ongoing dialogue with intended users, we can learn their reactions to results and prepare for future dialogues and for communicating results more effectively. If they find results surprising, and we are sure they are valid, we can learn more about why these results are surprising and think about ways we might enhance their understanding. Should we collect additional data? Invite others to participate in discussing the results? Communicate findings in more detail? Waiting until the end of the project to communicate results, in many settings, will prevent us from learning the reactions of our intended users and considering ways to maximize their use of the results. Similarly, delaying presentation of the results until the end will not give intended users time to think about the findings, to dialogue with the evaluator and others, and to consider their validity and potential uses. Reporting is an ongoing process, requiring different forms and means of interaction. It is ultimately about creating understanding and use, which is unlikely to occur without frequent communication.

We noted in Chapter 1 that evaluations can have many different purposes, and the information produced can be put to very different uses. For example, formative evaluation information is typically used by those wanting to improve a program they are developing or operating, whereas summative evaluation information is usually used by funders and potential consumers of the program to consider a program's successes and failures and to determine whether to continue it.

Reporting, then, also involves considering the purposes of the evaluation. In some summative evaluations with distant audiences who will use the results to make decisions about program continuation or expansion, communication may be relatively infrequent. Evaluators may provide the primary audience with interim written reports and then schedule an oral meeting to review the results before submitting a final report. In formative evaluations, however, evaluators are in frequent contact with the primary intended users and, in addition to formal occasions for presenting results, will also be chatting about the evaluation with intended users during visits to the program.

In addition to considering summative and formative purposes as major categories of evaluation, evaluators should consider many other potential purposes for the results and develop their reporting strategies accordingly. Henry and Mark (2003), Preskill and Torres (1998), Chelimsky (2001, 2006), Patton (2008a), and others have all discussed the range of ways evaluation findings can be used. Among the purposes they identify are:

- Demonstrating accountability
- Assisting in making a decision
- Bringing an issue to the attention of others (agenda setting)
- Helping stakeholders elaborate or refine their opinion of an issue
- Convincing others to take action
- Exploring and investigating problems
- Involving stakeholders in program planning or policy development
- Promoting understanding of issues
- Changing attitudes
- Changing the nature of dialogue or interaction among groups
- Influencing policy
- Introducing those involved to new ways of thinking through evaluation

Indeed, evaluation reports serve many purposes. Central to all of them, however, is that of delivering the message—informing the appropriate audience(s) about the evaluation, its findings, and conclusions in ways that result in understanding the results and their potential uses.

Different Ways of Reporting

Evaluators should first think broadly about the different methods they can use to report results. The method of reporting, as we will discuss, should be tailored to the audience. The most effective ways to catch the attention and interest of those audiences and stimulate their understanding and use of the results will differ, often dramatically. Torres, Preskill, and Piontek (2005) argue that learning is increased if results are communicated in more interactive ways—ways that encourage the audience to share their reactions and actively consider the results. In addition, of course, the evaluator learns more about their reactions. In Figure 17.1 we present

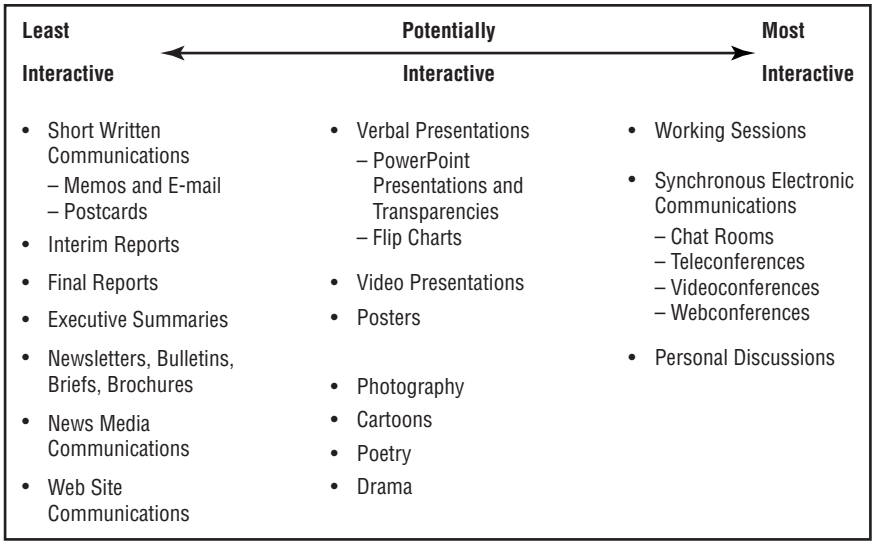


FIGURE 17.1 ***Communicating and Reporting Formats by Degree of Interaction with Audience***

Source: Evaluation Strategies for Communicating and Reporting: Enhancing Learning in Organizations by Torres, Preskill & Piontek. Copyright 2005 by Sage Publications. Reproduced with permission of Sage Publications.

their list of possible formats for communicating and reporting, categorized by level of interaction. Many of these are traditional, such as interim and final reports, but effective at reaching many audiences. Others are used frequently and should be considered for many audiences. They include newsletters, bulletins, and brochures; posters; and video presentations that might be posted on a web site. The list also emphasizes formats that we may neglect to use, such as working sessions, personal discussions, and teleconferences. Finally, formats that use photography, cartoons, poetry, and drama can capture the attention and interest of specific audiences and stimulate their understanding.

Important Factors in Planning Evaluation Reporting

In addition to purpose, factors to consider in evaluation reporting include:

- Accuracy, balance, and fairness
- The audience for the information
- When the information should be received or the most appropriate timing
- Effective communication styles

- Writing style
- Appearance of reports
- Sensitivity of the information
- Nature of the information to be communicated (positive, negative, neutral)

We will discuss each of these issues here and then address some of them in more detail later in the chapter.

Accuracy, Balance, and Fairness

It goes without saying that evaluation reports should not be unfair, unbalanced, or inaccurate. Yet truth is elusive, and even the most scrupulous evaluator must struggle to see that carefully collected and analyzed information is not later distorted, intentionally or unintentionally, in its presentation. As the Joint Committee (1994) states: "all acts, public pronouncements, and written reports of the evaluation [should] adhere strictly to a code of directness, openness, and completeness" (p. 109). Similarly, evaluators must be aware of their biases and reduce their influence on the presentation of information. Suppose the program director has been rude or often unavailable to the evaluator. These facts should not taint the judgments and language in the evaluation report, unless, of course, the program director also offends others in ways that have a negative effect on the program. Fairness in reporting is the hallmark of a professional evaluator.

Finally, there are two or more sides to every story. It is essential that legitimate positions be reported in a balanced way. No evaluator will ever be completely free of bias, but every effort should be made to control bias in reporting. The Joint Committee, evidencing its concern in this area, provided the following as one standard of accuracy:

> **A8:** *Communication and Reporting.* Evaluation communications should have adequate scope and guard against misconceptions, biases, distortions, and errors.

The American Evaluation Association's Guiding Principles also provide relevant guidance. They state that evaluators should

> **A3:** communicate the approaches, methods, and limitations of the evaluation accurately and in sufficient detail to allow others to understand, interpret, and critique their work. . . .
> **C5:** represent accurately their procedures, data, and findings, and attempt to prevent or correct misuse of their work by others. . . .
> **D4:** conduct the evaluation and communicate its results in a way that respects the stakeholders' dignity and self-worth. . . .
> **E3:** allow stakeholders access to, and actively disseminate, evaluative information and present evaluation results in understandable forms that respect people and honor promises of confidentiality. (http://www.eval.org/Publications/GuidingPrinciples.asp)

Tailoring Reports to Their Audience(s)

In Chapter 12, we discussed the importance of identifying the many stakeholders and audiences for an evaluation and suggested procedures for doing so. An evaluation report cannot be well targeted without clear definition of its audience and the types of questions that audience is likely to raise about findings. Writing an evaluation report before defining the audience is like firing a gun blindfolded, then hurrying to draw the bull's-eye in the path of the speeding bullet. As Lee and Holly (1978) note, "Identify your audience" may be an obvious, overworked platitude, but is an often overlooked step. They cite some common mistakes that have particular relevance for evaluation reports.

> Most evaluations have many audiences. Not identifying all of them is a common mistake. An ignored audience can on occasion get pretty testy and introduce a lot of undesired commotion into the situation. More typically, an audience who needs certain information, but never gets it, will make its decisions in ignorance of some perhaps vital information. . . .
>
> Another mistake you can make in identifying your audience is to identify too broad or too narrow an audience. An example of this would be for an evaluator to think a parent committee is the evaluation audience, when the actual audience is the committee chairperson. (She is the respected opinion leader of the group and always determines the action the committee will take on any issue.) Therefore, the majority of the evaluator's dissemination efforts toward the committee should be directed at informing and persuading the chairperson of the validity and implications of the evaluation information. (p. 2)

Different audiences have different informational needs. Knowledge of the values held by those who receive information can help the evaluator shape communications effectively. We suggest that an audience analysis be completed for all pertinent stakeholders. Such an analysis would involve determining what information each audience should be receiving or is interested in receiving, the values of the audience, and the best format or means for transmitting such information.

For example, methodologically oriented stakeholders, audiences, or colleagues will be interested in a complete, detailed report of the data-collection procedures, analysis techniques, and the like. Not so for the typical policymaker, manager, client, or public interest group. Neither school superintendents nor hospital staff will be interested in wading through descriptions of an evaluation's methodology. These audiences do not necessarily share the evaluator's grasp of or interest in technical details. Reports for such groups should be tailored so that the information of interest is conveyed clearly and in a way that establishes its credibility. The information, language, and level of technical detail should be appropriate to the audience.

Tailoring Report Content, Format, and Style to the Audience(s). Because of their diverse backgrounds, interests, preferences, and motivations, those who receive and use evaluation reports look for different things. A little reflection, and some

conversations with audiences, can help the evaluator identify the information of interest to each group. These actions begin your audience analysis. Observe what interests each audience most in meetings and personal interactions.

Although information needs will differ with each evaluation and each stakeholder group, in general program managers and staff are interested in the most detail on program operations, outputs, and outcomes. They know the program well and want to improve it. Stakeholders who are upstream from the program—administrators in the same organization, such as funders or policymakers external to the organization who may vote on issues relevant to the program—are typically more interested in information about outcomes and impact, though some such audiences are interested in client, parent, or citizen satisfaction with the program as well. The final important category of stakeholders, those who receive services and their families, are interested in program success, but generally in less detail than those overseeing the program. Those less familiar with the program or its totality (family members of clients or students) are also interested in brief program descriptions. (Here is an area where photographs can be useful in conveying program events.) Finally, audiences who are operating similar programs or serving the clients in other ways may be interested in more detailed program descriptions with a summary of program achievements. Of course, these suggestions are generic, but are intended to give the reader a sense for how audiences might differ in their information needs and interests.

In addition to including the specific content important to each audience, the evaluator must also account for differences in the ways audiences interpret and accept evaluation reports. One group may find inferences drawn from certain information credible and useful, whereas another group may scoff at the same conclusions (no matter how "scientifically" defensible). In evaluating a school program, testimonials of students and teachers might be the most persuasive information possible for some audiences, whereas others would prefer statistical summaries of student test performance. The evaluator must also take into account the criteria that various audiences will use to make judgments and the standards they will employ to determine the success or failure of that which is evaluated. What level of achievement constitutes success for each group?

Torres, Preskill, and Piontek (2005) discuss some of the major ways audiences may differ that should be considered in selecting a format for conveying information and developing that communication strategy. Those characteristics include:

- Accessibility
- Reading ability
- Familiarity with the program or evaluand
- Attitude toward and interest level in the program
- Role in decision making
- Familiarity with research and evaluation methods in general
- Attitude toward and interest level in the evaluation
- Experience using evaluation findings (p. 17).

The most accessible audiences are, of course, those groups or individuals that the evaluator sees frequently or at least occasionally in person. For an internal evaluator, or an external evaluator working closely with a program, the manager and the staff delivering the program may be quite accessible. In this case, personal conversations and work sessions provide an excellent forum for frequent communication. Less accessible audiences, which can include both upstream audiences and families of clients, need to be reached in other ways. Evaluators might communicate findings to parents on the evaluation of a school-based program in school newsletters and in talks with the PTSA or other parent groups. Upstream audiences who are intended users may serve on an advisory committee and learn results in this way. Advisory committee members serve as an important conduit for disseminating information on the evaluation and its results to the groups or individuals they represent. Encourage them to do so. Provide them with summaries or PowerPoints and, of course, volunteer to speak to their groups. But consider whether the group will receive the results best when presented by their representative or by the evaluator. Each has advantages, but it depends on the context, the nature of the group and its representative (how accurately will the results be portrayed?), and the evaluator (how sensitively and appropriately will the results be conveyed to the group?).

Using Technology. The advent of e-mail and the common use of the Internet have added an entirely new dimension to evaluation reporting. Most evaluation reports are posted on organizations' web sites, and such public access should be encouraged. However, the report should be developed with such access in mind. For example, the Internet and its technology can be used to incorporate pictures or videos of the program, which can enhance understanding and stimulate interest. Different versions (in length and topic) can be posted, allowing users to open the version of interest to them.

Today's blogs, twittering, and other electronic means of communication provide creative ways for communicating information and establishing dialogue with others. The obvious advantages of e-mail include its potential for instant and frequent communication between individuals or among members of a group. Its capacity for ongoing dialogues and its flexibility make it a prime medium, not only for routine evaluation reporting especially of interim reports and preliminary drafts of final reports, but also for atypical reporting. For example, evaluators can send preliminary findings and conclusions to the client(s) in bite-size segments, asking for their prompt reactions. Thus, clients can be involved in how evaluation results will be used. Google Sites can be used to develop web sites for advisory council members, evaluation team members, or primary intended users to share information on the evaluation. David Fetterman (Fetterman, 2001; Fitzpatrick & Fetterman, 2000) has made use of the Internet to share information, (field notes, photographs, quantitative data) to keep members of evaluation teams updated with evaluation findings and to continue a dialogue among the team members concerning their activities, discoveries, and interpretations. He posts filmstrips of evaluation sites on the

Internet so that any user of his evaluation findings can gain a picture of the setting and context for the study.

Audiences Can Help Tailor Reports to Fit Their Needs. Patton (2008a) points out that evaluation data are used more if the evaluator discusses and negotiates the format, style, and organization of reports with primary users. Brinkerhoff and his colleagues (1983) suggest that audiences indicate the information that is needed, the dates when it is needed, and possible displays or graphs that would be useful. We often share with clients mock graphs or tables (before data have come in) to focus their attention on the information they will receive and how it might be presented. Clients can then begin to understand the process and respond to different potential formats.

As we indicated at the beginning of this chapter, reporting is an ongoing strategy. Today, when many evaluators work to involve primary intended users and other stakeholders, ongoing communication is a way to improve their understanding, interest, and ownership of the results. At the same time, the evaluator can learn more from these stakeholders about how to communicate results to others like them or to audiences they know well.

Timing of Evaluation Reports

As purposes and audiences for evaluation reports vary, so will the timing of those reports. Formative evaluation reports designed to inform program administrators of needed improvements in a developing pilot program obviously cannot be delivered after the program has been completed (although that might be appropriate for a summative evaluation report to the program's sponsors or regulatory agency). An evaluation report that is limited in scope and that is perhaps even in rough draft form, but presented prior to relevant decisions, is preferable to a polished, comprehensive report that is delivered after those decisions have been made. Informal verbal briefings that serve an early warning function are preferable to formal but tardy reports. Timeliness is critical in evaluation.

The scheduling of evaluation findings must be guided in a general way by the role of the study. For example, early reporting will be more customary in a formative evaluation than in a summative study. But it would be wrong to conclude that summative reporting is restricted to formal, written reports distributed at the conclusion of the study. Indeed, too much formality may lessen the likelihood that evaluation findings will be used, for an evaluation's primary audience often will not take the time to study a report. Higher-level administrators and policymakers often hear evaluation findings only from their staff or others who have read the report and distilled from it the particular message they prefer. Evaluators who wish their message to be heard by managers should rely largely on informal interim reports, using nonprint strategies such as these:

- Being around and available to provide information that managers request
- Talking to those trusted people on whom the manager relies

- Using examples, stories, and anecdotes to make succinct, memorable points
- Talking often, but briefly, and in the audience's language

These suggestions are compatible with Cousins and Leithwood's (1986) report that the use of evaluation results is enhanced by ongoing communication and/or close geographical proximity between evaluator and decision maker.

Scheduled Interim Reports. Throughout the planning and implementation of the evaluation, the evaluator should schedule times to meet with stakeholders to share results and seek reactions. Remember that stakeholders are most likely to use findings that conform, generally, with their own perceptions (Weiss & Bucuvalas, 1980a). When findings are counter to potential users' conceptions or values, discussing interim findings with these stakeholders during periodic, regularly scheduled meetings can be helpful in preparing them for final results. Such meetings provide opportunities to explore their perceptions of the findings, change attitudes, increase the credibility of the evaluation and the evaluator, and, ultimately, increase the influence of the evaluation.

Reports can be scheduled at milestones in either the evaluation (e.g., conclusion of interviews, completion of data analysis on tests) or the program (e.g., near the end of budget cycles, semesters, or program cycles) or at regular intervals corresponding to routine meetings of clients or stakeholders (e.g., PTSA meetings, staff meetings). Internal evaluators have an advantage, as they are likely to be present or aware of occasions when results might be useful, but all evaluators should be alert to such occasions.

Unscheduled Interim Reports. The need for interim evaluation reports cannot always be seen in advance. No matter how carefully interim reports have been scheduled, there will be additional times when available evaluation information should be shared. In a formative evaluation, for example, the evaluator may discover a major problem or impediment. For example, she may discover that the video monitors used in an experimental program designed to train federal meat inspectors are too small for trainees beyond the third row to see the critical indicators of possible contamination. It would be a gross disservice to withhold that information until the next scheduled interim report, which might be weeks away, and then deliver the not-too-surprising message that a majority of the new generation of meat inspectors did not seem to be learning much from the experimental program that would serve the cause of public health. Helpful evaluators will deliver unscheduled interim reports whenever unexpected events or results pop up. Of course, unscheduled sharing of evaluation information is not limited to formative evaluation; as noted earlier, summative evaluators who wish to see their results used by managers learn to be around to share the emerging results of the evaluation informally and frequently.

Final Reports. Final reports are so familiar as to require no further comment here except to repeat that (1) they may be incremental (i.e., a preliminary final report

released for review and reaction by stakeholders, followed by a later final report), and (2) they need not be written, depending on the desires of the client. Because most clients do still request a written final report, however, we devote a later section to that subject.

Strategies to Communicate and Persuade

Written evaluation reports are nearly as varied as those who write them, but the great majority share a common characteristic: They make tedious and tiresome reading. Indeed, their variety seems limited only by the number of ways that can be found to make written information boring. Many deserve Mark Twain's waggish description of a particular book: "chloroform in print." One sometimes wonders whether such dreadful dullness reflects a purposeful design to discourage readers. Not that all evaluation reports are awful. Now and then one appears that is both interesting and informative, both enlightening and entertaining, both comprehensive and captivating. But these, like other gems, are rare.

Communication plays an important role in all stages of evaluation. Good communication is essential if the evaluator is to understand the origins and context for an evaluation, elicit the evaluative questions and criteria from stakeholders, reach agreements with clients concerning the evaluation plan, deal with political and interpersonal aspects of evaluation studies, maintain rapport and protocol during data collection, and so on. But nowhere is clarity of communication more central than during reporting. The quality of that communication will determine whether the evaluator's message comes through as clear or garbled, interesting or boring, constructive or hostile, and credible or incredible! Evaluators are providing information on something quite important to most stakeholders—something they have a stake in and beliefs about. Therefore, the evaluator must be sensitive to people's perceptions and styles, and consider the ways in which communication can help achieve interest, understanding, belief, and use of the results.

Construed broadly, communication may be thought of as all the procedures one person uses to inform another. Presenting information that cannot be understood is simply poor communication (no matter how correct the information). Presenting statistical summaries to lay audiences who do not understand statistics is poor communication (or noncommunication), regardless of how well a more statistically oriented audience might receive the same information. It is equally foolish to summarize rich qualitative data in prose that is truly literary and erudite when the audience consists of relatively uneducated stakeholders whose vocabulary and reading ability are far below that used in the summary.

Good communicators encourage evaluators to think of presenting results as telling a story. House (1980) argues:

> Every evaluation must have a minimum degree of coherence. The minimum coherence is that the evaluation tell a story. . . .
>
> There are at least two conventional ways of telling the story. One way is to present the evaluator as a neutral, scientific observer. In this case, the story line is implied.

> It runs something like, 'I am a detached, neutral observer who has made measurements according to the canons of science and have found certain things to be so. The program was conducted as I have described it, and I have found the following outcomes. . . .' Usually the story line concludes that 'the program was implemented, and such and such were the results.' Actual description is often sparse. . . . The usual presentation is to describe the project or the goals of the project, the treatment, the results or effects, and the conclusions.
>
> The second major way of telling the story is for the evaluator to stand closer to the program, as reflected in the narrator's 'voice,' and to tell the story by describing the events in detail. To this end the evaluator may use emotionally charged language and a narrative presentation. The story may look like a newspaper report. (pp. 102–103)

The importance of telling a story through the evaluation cannot be overemphasized. Consider what story-telling style or means of communication will be most persuasive with each audience. Some audiences may be accustomed to a dry, professional presentation and find that a more passionate one detracts from the credibility of the information. Many, of course, will find the latter, narrative story to be more engaging, to surprise them by catching their interest and increasing their understanding. Whatever tactic is used, it is imperative that evaluators take pains to consider what the message is and what they want each audience to learn and to communicate that information in a style that engages the audience and is most likely to result in learning.

Technical Writing Style

Nothing is as tiresome as reading tedious, unnecessarily convoluted, imprecise, and sometimes inconvenient and awkward expression. (See what we mean?) Wouldn't it have been better if we had said simply, "Nothing is as tiresome as reading complicated writing"?

We offer these few rules for improving writing style in evaluation reports:

- *Avoid jargon.* If users have certain jargon, however, it can be important to use their terms for clarity and credibility to them.
- *Use simple, direct language.* Make certain that the level of language is appropriate for the audience; don't ramble.
- *Use examples, anecdotes, illustrations.* Don't forget, a picture is worth a thousand words—and don't be flippant by asking why we did not illustrate this text.
- *Use correct grammar and punctuation.* Spelling should also be appropriate for the country in which the report is to be used.
- *Avoid cluttering narrative with reference notes.* Yes, we know we have done just that, but this is a text and reference book, not an evaluation report, and you are not the typical evaluation audience.
- *Use language that is interesting, not dull.*

Appearance of the Report

It would be interesting to do a bit of free-response research to determine what the first descriptive word is that pops into people's minds when they hear "evaluation report." We cannot predict the most common responses, but we would be amazed if they included *attractive*, *handsome*, or *visually appealing*. Concern with aesthetics has historically been as common among evaluators as compassion has been among tax collectors. Most evaluators have been preoccupied with what the report said, not with how attractively the message was packaged.

Appearance counts, however, because it will often influence how (or whether) a document will be read. Market analysts and advertising specialists could provide much useful information to evaluators who want their products used (such as how long it takes the average administrator to transmit most items from the incoming mailbox to the delete file or to delete a long message or attachment from their e-mail).

Until a decade or so ago, many evaluators seemed reluctant to adopt the slick and glossy visual tricks of advertising. Now most evaluators are concerned with the cover appeal of their report, as well as the attractiveness of its contents. Many evaluators frequently produce high-quality brochures and reports of evaluation findings for their various publics. Much knowledge from the marketing, commercial art, and publishing fields can be tastefully applied to make evaluation reports more visually appealing and readable.

Consider using a professional print organization that can provide advice on printing, binding, covers, color, and so forth for a print report and brochures. Budget for quality as well as quantity. Do not feel the need to follow academic publication styles. Your audience is not an academic one! (Save that for journal publications on your evaluation. That is the best way to reach that professional audience.) Instead, make liberal use of space and headings, font styles and sizes to produce a report that looks well organized, clean, and professional—one in which readers can easily find certain sections or information. A nearly endless array of computer software packages now allows artistically impaired evaluators to develop attractive and sophisticated graphs and art. The changes introduced by home-publishing software, the various "paintbrush" programs, and affordable color printers have left evaluators with no excuse for dull and drab reports.

Visual displays should be used in evaluation reports wherever they would be helpful in telling (or better yet, showing) the story. Use photographs, cartoons, and other illustrations to illustrate program actions or concepts. Careful use of color can make an evaluation report more attractive, as well as more functional. When the executive summary appears as the first section in an evaluation report, print it on colored paper. This not only gives some visual appeal but also draws attention to the summary and makes it easy for the reader to locate it later. Consider printing appendices on yet another color. It will be easy to turn to, and the combination of colors with the predominantly white body of the report enhances the visual appeal of the whole. Use graphic artists and/or web designers to develop covers of reports, separately published executive summaries, and brochures.

Obviously not all written evaluation reports warrant preparing and printing a cover. A typed and stapled cover page will serve well in many formative evaluation studies, and possibly some summative evaluation studies as well, but a more attractive report cover may entice readers and suggest that the evaluator thought the information contained in the report was worthy of a professional presentation.

Human and Humane Considerations in Reporting Evaluation Findings

Many evaluators become so preoccupied with preparing and presenting their messages that they forget the impact that those messages will have. If an evaluation report labels the U.S. Coast Guard's new officer training curriculum as ill-conceived and its implementation as inadequate, the personal egos (and perhaps the professional reputations) of the curriculum designer and the trainer(s) implementing the program will not go unscathed. This doesn't mean that truth should be diluted to protect feelings, only that it should be communicated as carefully, sensitively, and professionally as possible. Beyond apparent idealistic reasons for protecting the rights and sensitivities of those on whom the evaluation might reflect, there are also some obvious (if you think about it for a moment) pragmatic reasons. For example, in many evaluations, results are reported directly to those responsible for planning or running the program. Evaluators who are far too wise to tell any mother that her baby is ugly may tactlessly tell an administrator and staff that the program to which they have devoted three years of their lives is a disaster. Not surprisingly, the program practitioners may exercise the limits of their ingenuity in seeking ways to discount both the evaluation and the evaluator. The opportunity for the message to be of use might be irretrievably lost.

The evaluator must take appropriate steps to protect the rights and sensitivities of all those involved in the evaluation. For the practicing evaluator, this means that raw technical facts must be told with sensitivity and in a context of trust. In this section we offer our suggestions for (1) delivering negative messages, and (2) providing those affected with the opportunity to review a draft report (and suggest modifications) prior to its final release.

Delivering Negative Messages

In olden days, the messenger who delivered news to the king lived a life fraught with risk. If the news was bad, the messenger might lose his tongue—or even his head. Nowadays, bearers of bad evaluation tidings may still find themselves savaged (though in a somewhat more polite manner).

Sometimes evaluation clients (or others involved with the evaluation) are so sensitive to any criticism or any hint of imperfection that it would not matter much how negative findings were reported—the reaction would still be defensive. But more often we have observed that defensive reactions are exacerbated by the manner in which the negative results are conveyed.

It is helpful for evaluators to begin with the strengths of the program. (To those who say they cannot find any strengths to report, we suggest they are not

very thorough or insightful; in even the most awful program, one can usually comment sincerely on the efforts, dedication, and hard work of the staff.) We also recommend the following steps suggested by Van Mondfrans (1985) for helping those involved in evaluation to swallow bitter pills:

1. In an oral debriefing where the major events of the evaluation are reviewed and where the major findings are previewed, the negative information is stated in as positive a context as possible. It seems easier for clients to accept negative information in an oral form in a relatively friendly encounter.

2. A preliminary written report is presented in which the negative information is described in a straightforward factual manner but from as positive a perspective as possible. Often a personal visit should follow in which the preliminary report is discussed and the client allowed to propose changes if the information is viewed as unfair. If changes are needed in the preliminary report, they should not obscure negative information or allow it to be misinterpreted; however, it may be that in discussing the negative information, the client will bring up other factors not known to the evaluator at the time the preliminary report was written. These other factors may be included in the final report in juxtaposition with the negative information, thus allowing better interpretation.

3. A final written report is prepared in which the negative information is accurately and fully presented. The client is better prepared to deal with the negative information having had a chance to review it several times, to think of other factors that are relevant, and present those to the evaluator. The evaluator has the opportunity to review other factors and include them in the report if they aid the interpretation of the negative information (pp. 3–4).

Benkofske (1994) has also suggested ways to prepare stakeholders to receive negative findings about their program's performance. In her view, this is especially important when the negative findings are distilled from qualitative data. She reports that most of her clients have heard so much about the benefits of qualitative data that they inappropriately come to expect it to yield glowing descriptions of their programs. Therefore, she finds it important to engage clients early on in a discussion of what they will do if the qualitative data turns out to be negative. She says:

> Clients all come believing their program needs a qualitative study. I have found, however, that stakeholders need to be prepared for the evaluation results if the qualitative data are not positive. It is my personal experience that qualitative data hurts; it stings like hell to see in print pages of quoted material that describes in vivid detail problems with a program. While qualitative data can "brighten the portrait" when positive, it can wound deeply when negative. (Benkofske, 1994, p. 2)

So, how does one prepare clients for negative findings?

1. *Prepare clients for some disappointing findings from the beginning.* At early meetings, it can be important to note that all evaluations find some successes and some

failure. Few programs achieve all their goals at the desired level. Remind key stakeholders of that often.

2. *Let key stakeholders know of negative findings promptly.* As evaluators encounter surprising and typically negative findings, mention them to the key stakeholders soon, perhaps in a general way, such as, "I was surprised at some of the interviews. People aren't really findings jobs like we thought they might." Observe their reaction. Then, move to discussing the results in working sessions either with individual stakeholders of concern or groups of those stakeholders. Seek their interpretations and, if formative, move to discussing potential solutions.

3. *Listen to stakeholders' concerns or questions about data accuracy.* When data, methodologies, or interpretations are questioned, evaluators should take those questions as seriously as they expect the audience to take the results of the evaluation. The evaluator should go back and examine the data in other ways or consider collecting additional data, of different types or from a different source, to attempt to triangulate the findings from different sources and add validity. Such actions should gain the concerned audiences' trust in the evaluation and in the evaluator. The evaluator is providing a role model for the behavior he or she hopes the client will demonstrate: questioning, seeking truth, being open to failure, learning, and looking for ways to improve.

4. *Move to considering actions resulting from disappointing findings.* As the client accepts the negative findings, evaluators should move to discussions about ways to present the information to others in oral presentations and final reports, as well as ways to improve the program in these problematic areas, if the purpose is formative. As we have noted, not all information has to go to every audience. The audience analysis will help evaluators consider who needs to know what. Thus, some negative information, and the details of it, need go only to limited audiences.

5. *Allow clients to review, and even make suggestions for, ways to present negative findings to others.* The evaluator should maintain the accuracy of the report, but should balance the means of communicating these findings with the extent to which the manner of communication will enhance or impede use of the results. If the intended users are the involved clients and they have ultimately proven to be responsive to the findings and suggestions for change, their input on the means for communicating the findings to others can be welcomed. If the intended user is a different audience—not the stakeholder(s) who have been most closely involved—the evaluator must begin to communicate findings to this stakeholder in the ways discussed previously.

Providing an Opportunity for Review of the Report. Only the most arrogant evaluators would assume that their work and the report that presents it are completely accurate and fair in all regards. Small factual errors can lead to nontrivial errors in judgments and conclusions. Interpretations can overlook contextual factors that the evaluator failed to appreciate and thus be incorrect. And the evaluator's bias can creep into evaluation narratives unnoticed, especially by the evaluator.

For all of these reasons, we strongly urge the evaluator to circulate a draft of the evaluation report to the client and other key stakeholders for comments, asking that they point out (and correct where appropriate) problems with wording, factual errors, or mistakes in ultimate interpretations or judgments.

Reviewers should be asked not only to challenge anything that they perceive to be in error but also to provide substantiation for the alternative facts or interpretations that they propose. Reviewers may need to be reminded that the evaluator is under no obligation to accept their suggestions (the intent is not to allow clients to rewrite the report any way they wish) but only to give those suggestions serious consideration. The evaluator reserves the right to ignore suggestions and to make only those changes that are warranted. (These issues—the rights of the evaluator to the wording of the final report, the rights of stakeholders or clients to input—should be clarified during the earlier, contracting stages of the evaluation. See Chapter 14 on evaluation agreements and contracts, and managing evaluations.)

Circulating a preliminary draft report can increase the number of individuals who read the report carefully; shared responsibility for the report's accuracy is a good motivator. Some may worry that the use of drafts may lessen interest in the final report. That concerns us less than the very real possibility that many key persons who are not asked to review a draft may never read the report at all. As we have stated, the final report should come as no surprise to the primary intended users or stakeholders. They should already have learned the information contained in it from the evaluators and be familiar with the details.

What if the evaluator refuses to accept a proposed change in the report, but the reviewer who suggested it continues to contend that the report is inaccurate, misleading, or unfair? Simple. Invite the reviewer to share that view, in writing, and include it as a minority report. We see no problem with permitting reviewers to include their rebuttals, rejoinders, or contrary comments. If the evaluator's data collection, analysis, interpretation, judgments, and conclusions are on solid ground, they should not be harmed by such detraction. If they are shaky and cannot withstand such challenge, then they deserve to be challenged.

Key Components of a Written Report

No one best outline or suggested table of contents fits all written reports. Evaluation roles, objects, and contexts are simply too diverse to permit that. Each one contains idiosyncrasies peculiar to itself, and reports must be tailored to reflect such uniqueness.

Yet there are some important items that should be included in almost every written evaluation (at least every formal, final evaluation report, and interim reports as appropriate). These items are the core of most good written evaluation reports.

We believe that one must be much more concerned about the form of formal reports intended for external audiences. We see the following outline as

applicable in other situations as well, however, and offer it as a heuristic checklist that evaluators might consider as they prepare any written evaluation report.

A written, comprehensive, technical evaluation report will typically contain the sections listed in the following generic table of contents:

- **I.** Executive summary
- **II.** Introduction to the report
  - **a.** Purpose of the evaluation
  - **b.** Audiences for the evaluation report
  - **c.** Limitations of the evaluation
  - **d.** Overview of report contents
- **III.** Focus of the evaluation
  - **a.** Description of the evaluation object
  - **b.** Evaluative questions used to focus the study
  - **c.** Information needed to complete the evaluation
- **IV.** Brief overview of evaluation plan and procedures
- **V.** Presentation of evaluation results
  - **a.** Summary of evaluation findings
  - **b.** Interpretation of evaluation findings
- **VI.** Conclusions and recommendations
  - **a.** Criteria and standards used to judge evaluation object
  - **b.** Judgments about evaluation object (strengths and weaknesses)
  - **c.** Recommendations
- **VII.** Minority reports or rejoinders (if any)
- **VIII.** Appendices
  - **a.** Description of evaluation plan/design, instruments, and data analysis and interpretation
  - **b.** Detailed tabulations or analyses of quantitative data, and transcripts or summaries of qualitative data
  - **c.** Other information, as necessary

A brief discussion of each of these major sections and their contents follows.

Executive Summary

One feature of many evaluation reports that makes them so formidable is their organization. It often requires the busy reader to ferret out from a compulsively detailed report why and how the study was conducted and what important information it yielded. Sometimes a brief summary of essential information is wedged somewhere between the presentation of the findings and the appendices, but often readers are left to sift out the most valuable nuggets of information for themselves.

Most evaluation audiences do not have (or will not take) the time or energy necessary to read a thick report laden with tabular information or narrative details. It makes good sense, therefore, to provide a brief executive summary in one of the following forms.

Executive Summary within a Report. For most evaluation studies, an executive summary might best be included within the report itself, preferably right up front so it is the first thing the busy administrator or provider sees when the report is opened. We also propose that the executive summary be printed on a different color paper to draw attention to it. This summary is usually between two and four pages in length, depending on the scope and complexity of the evaluation. In addition to a very brief description of the study's purpose, and a very brief word about how data were obtained (e.g., "Data were collected with questionnaires mailed to agency employees and a focus-group interview with agency managers"), the summary should contain the most important findings, judgments, and recommendations—perhaps organized in a simple question-and-answer format or with major findings or recommendations bulleted or numbered. If the evaluation report is large and interest in it is broad, then it is sometimes more economical to distribute a separately bound executive summary similar in all other respects to what we've just described.

Executive Abstract. With a large evaluation audience, it may be necessary to condense the executive summary to a one- or two-page abstract that contains only the major findings and recommendations without any supporting documentation. Such abstracts are often useful in communicating evaluation results to large legislative bodies, parents, citizens, community leaders, members of professional associations, and the like.

In one statewide evaluation of a controversial program conducted by one of the authors, three interrelated written evaluation reports were prepared: (1) a large, detailed technical report containing most of the information called for in the earlier outline; (2) a medium-size summary of major interpretations and judgments drawn from the data; and (3) a brief executive summary of the study purposes, findings, and conclusions. Availability of these three reports was broadly announced in the newspapers and on television. Readership was estimated by the number of people who requested a copy or checked one out in the several repositories in which they were made available. Nearly 400 individuals read the executive summary, 40 read the midsize interpretive report, and only one person even requested the complete report (and he was an expert methodologist hired by opponents of the evaluation to see if he could find fault with it). As these results show, shorter reports will often be most widely disseminated.

Introduction to the Report

Despite its prominent placement in the report, the executive summary is only a brief abstract, not an introduction. An adequate introduction will set the stage for the remainder of the report by outlining the basic purpose(s) of the evaluation and the audiences the report is intended to serve. For example, is it intended to provide information to legislative budget analysts who will make recommendations for the future funding of new forms of performance evaluation for teachers, or is it to

document the performance of Mid-City's implementation of the state teacher evaluation program in a field test? Is the audience the state legislature, those who developed the program at the state department of education, the administrators and staff in individual school districts operating the program, or just the staff at Mid-City?

One good way to ensure that a report will be relevant is to describe thoroughly the rationale for the evaluation. The rationale should address such questions as these: Why was the evaluation conducted? What is the evaluation intended to accomplish? What questions was it intended to answer? Why was the evaluation conducted the way it was? Once this information is provided, audiences can determine whether the report is relevant by asking how well each question is answered.

The introduction is also one logical place to caution the reader about limitations that affect the collection, analysis, or interpretation of information. Such limitations should be openly disclosed here (or in a later section dealing with evaluation procedures). Similarly, disclaimers are sometimes placed at the beginning of a report (e.g., in the preface or on the title page) to clarify what the evaluation is and is not, thus protecting both clients and evaluators from criticisms based on misunderstandings.

It is also useful to provide a brief reader's guide to the report in the introduction. The table of contents only lists major topics. The reader's guide explains what each topic comprises.

Focus of the Evaluation

This section describes the program to be evaluated and the questions the evaluation will answer. Although a full program description, as discussed in Chapter 12, will rarely be provided here, a shorter program description is important for those who are less familiar with it. This section might present a brief history of the program (when and why it began and who initiated it), a logic model and the program theory with narratives discussing the critical parts of each, program goals and intended outcomes, a description of the staffing of the program, the numbers of clients served and their characteristics, and any important contextual issues—for example, location, oversight, legislation, or regulations. Some of these descriptive elements (logic model, program theory, description of staffing, etc.) may be included in an appendix if they are too detailed for this section.

It is also important to list in an early section the evaluative questions used to focus the evaluation. If differential priorities were assigned to the questions, that process should be explained.

Finally, it is useful to include a subsection outlining the needed information the evaluation was intended to collect, analyze, and report. Such a list helps make the rationale for the next section much more apparent.

Brief Overview of the Evaluation Plan and Procedures

Any complete evaluation report must include a detailed presentation of the evaluation plan, the data collection instruments, and the methods and techniques used to analyze and interpret data, but these need not be in the body of the report.

Early in our evaluation careers, we included all such details in this section. After a decade or two, we each concluded that a careful and complete summary would suffice in this section if detailed procedures (and possibly even the instruments themselves) were contained in supporting appendices. After another decade, we each moved the entire description of the design, instruments, data collection, and analysis procedures to an appendix, leaving in this section of the report only a few sentences that tell very generally where the data came from and how they were obtained. Readers of the report wishing more detail are referred to the appendix.

Presentation of Evaluation Results

This section of the report contains the results of the evaluation and represents the source of subsequent conclusions and recommendations, preferably in the form of a complete summary, using tables, displays, and quotations as appropriate, and referencing more detailed data summaries or transcripts in supporting appendices. Some audiences are put off by too much statistical data (and factor analyses, multiple regressions, and the like should be avoided in the narrative with almost any nontechnical audience). However, many policymakers, managers, and others respond positively to data presented in straightforward graphs and charts because they are able to summarize data in a way that many users can understand (Alkin, Stecher, & Geiger, 1982; Henry, 1997). Henry (1997) provides a sourcebook on ways to present graphs effectively.

Remember, too, that numbers generally fail to adequately portray or illustrate the program and its impact on others. Quotations from interviews with clients or community members, pictures of program activities, and mini-case studies or stories of individual students or other recipients of services can be quite effective in giving readers a deeper understanding of the issues. (See Fischer and Wertz [2002] for a discussion of four formats they used to convey results of their study of victims of crime and their rehabilitation to policymakers.)

The interpretation of the results is as important as their presentation. Evaluation depends, after all, on the evaluator's ability to perceive and interpret. Interpreting data should not be an informal or casual activity. Rather, it should be a careful process, made as public as possible by the evaluator's careful listing of all content and steps followed to reach the particular judgments and recommendations presented.

One of the most disconcerting deficits of many evaluation reports is the lack of any organization that will assist the reader in relating findings to the major evaluative questions posed. Without organization or categorization, findings often blur, becoming less understandable. We urge evaluators to relate their findings to the most logical set of organizers. We prefer to organize the findings around the evaluative questions posed for the study in a question-and-answer format. Other organizers might include the goals or objectives (if that is the evaluation's focus), various components of the program, or different groups of clients. Whatever the organizer, some structure must be given to all but the most simplistic, single-variable, single-question evaluations.

Conclusions and Recommendations

In this section of the report, we first summarize the major conclusions of the evaluation. Often, the results are presented in detail and the reader has lost the big picture. In other cases, readers will skip the results and move to the conclusion. Therefore, beginning this section with a listing of major findings, perhaps organized by evaluation question, is essential.

The report can then move to a synthesis and discussion of these findings. What is their meaning? How do they correspond to standards or criteria that may have been established at the beginning of the evaluation to judge the program?

Standards and criteria should be listed explicitly. The data do not speak for themselves. The evaluator who knows the data well is in the best position to apply the standards to the data to reach a judgment of whether the evaluation object is effective or ineffective, valuable or worthless. Making judgments is an essential part of the evaluator's job. An evaluation without clear criteria is as much an indictment of its author's lack of sophistication as one in which judgments are not based on the data.

We strongly prefer organizing evaluative judgments under the headings *strengths* (presented first) and *limitations* (or the parallel and more familiar *strengths* and *weaknesses*, if client and evaluator are less squeamish). Several advantages accrue to this dichotomous presentation:

- Attention is focused on both positive and negative judgments.
- Audiences can conveniently locate the evaluator's positive or negative judgments.
- Presenting strengths first generally helps those responsible for the object of the evaluation to accept the weaknesses listed thereafter.

The discussion of strengths and limitations must be complete enough to allow the audience(s) to see the rationale and judgments on which later recommendations are based. Another useful format that is familiar to planners in corporate and higher education settings is the SWOT format (strengths, weaknesses, opportunities, and threats).

We prefer the report to conclude with a series of recommendations. These recommendations can be summative in nature, if that is the purpose of the evaluation. They should directly address whether the program should be continued and, if so, whether it should be expanded to other sites and the nature of those sites. If the recommendation is to discontinue, the evaluator might make recommendations for other interventions that could be considered if the need continues. If the program is to continue, the report typically contains some formative recommendations for program revision and improvement in its continuation. Formative evaluations, of course, contain recommendations solely for program improvement. Many detailed recommendations concerning implementation may have already been made, through other means or other reports, to program staff and managers.

This final report is not the place for detailed recommendations. Instead, its purpose is to let others know the general nature of these recommendations, or perhaps to hold managers and staff accountable for making the suggested changes or to permit upstream stakeholders to ask more questions about the recommendations and their cost and rationale.

Often, evaluators think they do not know enough to make specific recommendations—that their job is to collect data and to make a judgment based on it. But, the actions that should emerge from the report's conclusions and judgments are often not immediately obvious to readers. Just as some formative recommendations may have been made in other forms to program staff and managers, other recommendations may be discussed in detail with funders or administrators of the program. However, again, the report should provide a summary of these recommendations, but the evaluator should make some recommendations to encourage action. In some cases, when the evaluator lacks sufficient knowledge for action, her recommendations may urge that attention be given to correcting a problem without specifying the exact means by which the problem should be corrected. In other cases, it can be very useful for the evaluator to work with the advisory group or, if none exists, the client, and possibly other stakeholders, to develop recommendations that do propose feasible and appropriate corrective actions. Such recommendations, developed in collaboration with stakeholders, can help increase confidence in the report.

There are times when recommendations might be appropriately omitted. In some cases, a report may be used to begin a strategic planning process to generate recommendations. The evaluator might serve as the facilitator of that process or, if he lacks skills in strategic planning, he should certainly be present as a resource. But in this case, a mechanism has been established to generate recommendations.

Minority Reports or Rejoinders

As discussed earlier, it is sometimes important to include a section in which those who disagree with the evaluator's judgments, conclusions, or recommendations can share their dissenting views. Or, if one member of an evaluation team disagrees with the majority view, it seems sensible to insert any rebuttals or minority reports as a last section.

Appendices

Supporting appendices (bound within the report or as a separate volume) contain information that is needed to help the reader understand such things as what (if any) sampling procedures were used, how the information was collected to ensure its accuracy, and what specific statistical or narrative analysis procedure was used and why. In short, this is where the evaluator provides the methodological and technical information needed, not only by primary audiences, but also by fellow evaluators who will decide whether the conduct of the study was adequate to make its results believable. The evaluator should not forget that fellow evaluators

who are keenly interested in methodological and technical adequacy may be perusing those reports. It is wise to remember Campbell's (1984) insistence on "having available (along with the data available for reanalysis) a full academic analysis for cross-examination by our applied social science colleagues" (p. 41). The appendix is the best place for such detailed descriptions of evaluation procedures, data tabulations or analyses, observation logs, complete transcripts of important interviews, and other information that is relevant but too detailed to present in the body of the report. Appendices might also include the actual data-collection instruments and any other information (e.g., boundary maps of sampling units in a community survey) that is interesting and important to the audiences but inappropriately detailed and/or too extensive for inclusion in the body of the report. Appropriate use of appendices will make the report itself much more streamlined and eminently more readable.

Suggestions for Effective Oral Reporting

Written evaluation reports, although very common, are not always the most effective medium for reporting. Oral reports, supported by appropriate visual aids, can be more effective at catching interest, stimulating dialogue and interaction, and ensuring understanding. They provide an opportunity for eye contact and, thus, for gaining a sense for others' reactions even before the end of the presentation. Some oral presentations provide opportunities for interaction during the presentation, which can be an effective way of engaging the audience, providing a break and change of pace for a longish presentation, and learning about the audiences' views. Finally, oral presentations provide the opportunity for questions and answers and for obtaining comments, suggestions, and ideas. None of these are possible with a written report, although distribution of drafts through e-mail or on a web site can simulate the oral process. However, the focus remains on the written report.

Many of the earlier suggestions for improving written reports are pertinent for oral reports as well. Consider the following questions: Who is the audience? Is the report to be presented in a small meeting with organizational staff, managers, or funders? Or is the presentation to a larger group of people who deliver the program, clients and/or their families, or members of the community? What is the nature of the forum? Is it formal or informal? How much time is provided for the presentation? Does the evaluator know the audience personally or are they strangers, new to the evaluation? All of these are critical issues to consider in adapting the oral presentation to the interests and information needs of the audience, and preparing it in such a way that it conveys the necessary information to the audience and the evaluator. The audience learns some issues of interest to them about the program. The evaluator learns what this audience thinks of particular aspects of the evaluation and the program.

Just as short memos on the evaluation differ dramatically from the final report, so oral presentations can range from a few minutes at a staff meeting to a formal presentation before a board of directors or a school board. Thus, it is difficult to generalize in our advice on oral presentations, but we will highlight a few considerations.

The most important element is, again, tailoring the presentation to the audience. Learn about the audiences' norms and expectations. Are they formal or informal? What is the audiences' primary interest in the program and in the evaluation? What do they already know about the program and the evaluation? What do you want from them (reaction, perceptions, different uses of the results)? As with a written report, evaluators should consider the following principles:

- Accuracy, balance, and fairness
- Communication and persuasion
- Level of detail required (an oral presentation is not the place for providing great detail unless it is requested as a relatively long, formal presentation)
- Use of simple, direct, correct, and interesting language
- Avoidance of jargon and unnecessary technical language
- Use of examples, anecdotes, and illustrations
- Sensitivity to the rights and feelings of those involved

Oral reports also require particular attention to audiovisual presentation of information. Obviously the suggestions commonly offered in speech and communications courses and texts are relevant here, but the following tips are particularly relevant for making effective oral evaluation reports:

1. *Determine the story you want to tell.* What information is necessary to communicate that? What anecdotes or personal stories of the program might be useful to illustrate key points? How can visuals be used effectively (pictures, cartoons, tables/graphs, flowcharts, bulleted points)?

2. *Decide who should tell the story.* It is not essential that the lead evaluator also be the lead storyteller; what is essential is that the story be told well. If the lead evaluator has that capability, then that person would obviously be the best choice. But using another member of the evaluation team (or even an outside "reporter") is far preferable to having a good evaluation destroyed in its telling by an evaluator who is not a strong presenter. Usually the lead evaluator(s) can be involved in at least part of the presentation so that any awkwardness can be minimized. Consider whether the presenter has competence in the culture of the audience. This can be demonstrated by use of language, familiarity with customs, geography, history, and so forth.

3. *For formal presentations, select the oral report medium* (verbal narrative, videotape, staged debate, presentations by clients or students, etc.). Make the presentation format interesting and varied using multiple media, multiple presenters, or other variations. Don't use the format the audience expects; do something different to gain and maintain interest.

4. *Make visuals to accompany the presentation.* But notice the word "accompany." Visuals should not dominate or lead the presentation. Many otherwise effective PowerPoint presentations are ruined by their being permitted to dominate the presentation. Reading a list of bullets from a PowerPoint presentation is not an effective use of visuals. Conversely, using PowerPoint presentations to highlight major issues; to present creative graphics, pictures, images, flowcharts, or complex tables; or to inject humor and color into a presentation can awaken the audience to the possibilities of the findings. Make sure to test the presentation at the place where it will be delivered. More than one presentation has failed because the evaluator did not determine if the computer capabilities at the site matched those anticipated for the presentation.

5. *Develop a presentation that feels natural and comfortable to you,* then practice until you are at ease delivering it. Use effective highlighting techniques, such as a laser pointer. Practice using it. You don't want your laser beam to hit a client in the eye!

6. *Involve the audience in the presentation.* During the presentation, you might ask for a show of hands on different issues or experiences. Or, you might give them three minutes to talk in groups of two or three about their thoughts and recommendations. Be sure to leave plenty of time for questions. If the presentation is long, stop at certain points to take questions.

7. *Develop and adhere to an agenda.* Most effective oral presentations are relatively short, but, as with the formality of the style, the length depends on the purpose. Many presentations are five to ten minutes long. Fifteen minutes is too long for a community audience and even for most stakeholders. Protracted oral reports will have people slumping in their seats (or worse, out of their seats). Use handouts of executive summaries or brochures designed specifically for the audience to provide additional information, and use the time of the oral presentation to capture and stimulate interest and questions and to convey major points. Oral presentations are intended for interaction. Allot time for questions and discussion. In most cases, the time for Q & A should be longer than the actual presentation. Remember, the purpose is not just for the evaluator to convey information but for the evaluator to learn from the audience.

Of course, *frequent* oral reporting of results—in staff meetings, forums with clients or community members, individual meetings with key managers or policymakers—is the critical way to engage audiences in the evaluation process and increase its ultimate influence. Evaluators should attempt to attend many meetings with such groups, taking the opportunity to observe the types of information each group expects, their style of interaction, and their information needs. In addition, evaluators can use the meeting to insert a few key pieces of information being gained from the evaluation or to remind audiences of the progress and thank them for their assistance. These frequent, more informal methods of communication prepare the audience to receive the final report with curiosity, interest, and optimism.

A Checklist for Good Evaluation Reports

Ingredients of a good evaluation report can readily be inferred from our earlier suggestions, but here, for convenience, is a checklist of things that would typify most good evaluation reports.

_____ Interim and final reports provided in time to be most useful
_____ Report content, format, and style tailored to the audience(s)
_____ Involvement of audiences in determining the format and style of the report
_____ An executive summary
_____ An adequate introduction to set the stage
_____ Mention of limitations of the study
_____ Adequate presentation of evaluation plan and procedures (primarily in appendices)
_____ Effectively organized presentation of results
_____ All necessary technical information provided (preferably in appendices)
_____ Specification of standards and criteria for evaluative judgments
_____ Evaluative judgments
_____ Lists of identified strengths and weaknesses
_____ Recommendations for action
_____ Protection of clients' and stakeholders' interests
_____ Sensitivity to those affected by the evaluation findings
_____ Provision for minority reports or rejoinders
_____ Accurate and unbiased presentation
_____ Effective communication and persuasion through telling the story
_____ Appropriate level of detail
_____ Lack of technical jargon
_____ Use of correct, uncomplicated, and interesting language
_____ Use of examples and illustrations
_____ Attention to visual appearance and eye appeal

How Evaluation Information Is Used

The utility of any evaluation is a prime criterion for judging its worth (Joint Committee, 2010). Use is one of the factors that helps distinguish evaluation from research. Evaluation is intended to have an immediate or at least a near-term

impact; while research is intended to add to knowledge and theory in a field, the results it yields may not be used for some time. If an evaluation study appears to have no effect, it will be judged harshly regardless of its technical, practical, and ethical merits, whereas research is not held to the same criterion of usefulness.

Evaluators have been studying and writing about evaluation's use since the early years of the profession (Suchman, 1967; Weiss, 1972). Many respected evaluators in the 1970s and early 1980s reported that evaluation results were often disregarded (Cousins & Leithwood, 1986; Patton, 1986; Weiss, 1977). However, more recently, observers have suggested that these earlier commentators underestimated the actual impact of evaluation studies. For example, Cook wrote, "in the past decade it has become clearer that instrumental [direct] use does occur and that prior accounts of its demise were exaggerated" (1997, p. 41).

Evaluators have traditionally identified three types of use for evaluation:

- Instrumental use in which findings from an evaluation are used directly to make changes in a program (formative) or its funding or continuation (summative);
- Conceptual use in which the findings of an evaluation provide new information (new concepts) to users, but that information does not result in action or use though it may change attitudes or beliefs about elements of a program;
- Symbolic use in which evaluation results are used in symbolic ways to continue current or predetermined actions (Leviton & Hughes, 1981).

In the last few decades, evaluators have recognized that other uses take place. Patton (1997c) first defined the term "process use," which Cousins (2003) distinguishes from the previous emphasis on instrumental use of evaluative findings. Process use occurs as a result of being involved in an evaluation because those activities, that participation itself, can result in learning. Patton defines process use as "individual changes in thinking and behavior, and program or organizational changes in procedures and culture, that occur among those involved in evaluation as a result of the learning that occurs during the evaluation process" (1997c, p. 90). These changes can influence the development of future programs or create an openness to new ideas for the organization or new ways of thinking.

The concept of process use has prompted many evaluators to focus their efforts on organizational and individual learning. As Cousins and Shula observed, "Evaluation from this perspective is less about practical problem-solving and more about learning" (2006, p. 271). Several prominent evaluators have conceptualized program evaluation as an intervention in the organization to influence organizational behavior or policy (Cousins, 2003; Henry & Mark, 2003). Work by Preskill and Torres (1998, 2000) and Preskill's 2008 presidential address at the American Evaluation Association focused on the role of evaluation in working with organizations to enhance learning. (Torres, Preskill, and Piontek [2005] cited earlier in the reporting section, is designed to focus reporting on enhancing learning through evaluation.) Evaluation capacity building (ECB) also emerged from this focus; the

impact of evaluation could be on changing how an organization operated and how those in the organization thought about programs.

Another recent emphasis in understanding evaluation use has been on context. Use does not occur in the same way in every setting. Instead, the context for both the program and the organization influences the ways in which evaluation is used. Political pressures can enhance or impede the use of evaluation. Some programs that are found to be ineffective continue because of strong stakeholder or public support. Evaluators have learned that, aside from political influences, organizations operate in nonrational ways. An organization's culture influences its openness to new or different ideas.

Finally, individual users and their characteristics influence evaluation use. One of the first studies to find this impact was a study by Weiss and Bucuvalas (1980a; Weiss, 1983), which remains well known. They first assessed mental health professionals' attitudes toward an issue, then sent them an evaluation report on the issue and asked them to give feedback on the report. Their analyses revealed that those administrators whose views coincided with the findings of the report were much more likely to view the report as valuable. Those whose views, based on their own knowledge and experience, conflicted with the report's findings were critical of the study's methodology and rejected the findings. This and other research prompted Weiss and Bucuvalas to coin the terms (and title of their book) "truth tests" and "utility tests" as the means by which evaluation reports were judged. A report passed the truth test if the results corresponded with the readers' own knowledge and experience. It passed the utility test if its results were perceived as useful. Cousins and Shula (2006) draw on research concerning knowledge utilization to help consider other individual factors that influence evaluation use. Among the pertinent findings for evaluation: information is best received when the giver and receiver share similar status and beliefs (Rogers, 1995), and people are more likely to use information if it serves their self-interest in terms of contributing "directly to some economic gain, social prestige, work convenience or sense of satisfaction" (Cousins & Shula, 2006, p. 274; Rich & Oh, 2000).

Models of Use

Kirkhart (2000) and Henry and Mark (2003) have developed models that help us to extend and define the types and nature of effects that evaluation might have in different settings. Kirkhart, building on Patton (1986) and others (Weiss, 1980), emphasizes that the language we use affects how we conceptualize evaluation use. She advocates the use of the term "influence" ("the capacity or power of persons or things to produce effects on others by intangible or indirect means") to convey evaluation's many potential effects rather than the narrower term "use" (2000, p. 7). She also proposes a model of an integrated theory of influence to depict the different potential effects of evaluation. (See Figure 17.2.) Her model highlights three dimensions: the source of the influence, the intention, and the timeframe.

The *source of influence* dimension makes us aware that evaluation can have an influence through either its results or its process. As we noted previously,

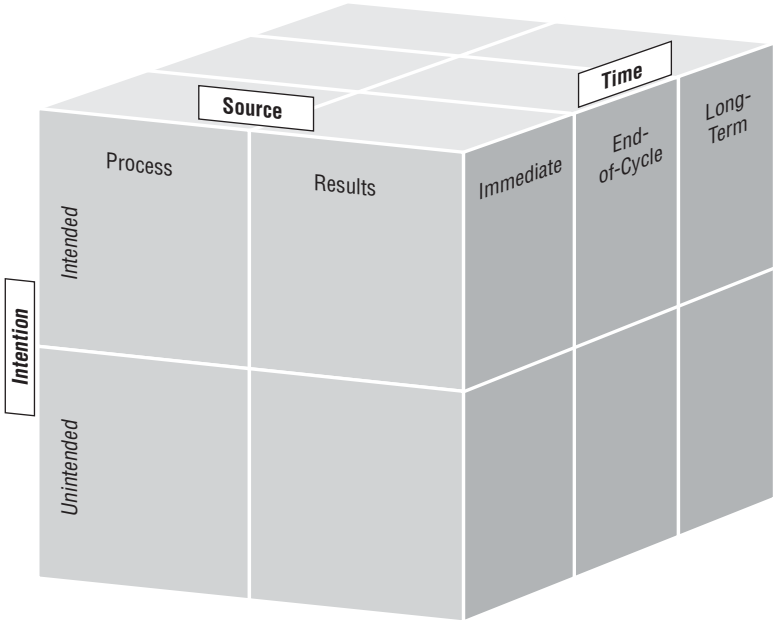


FIGURE 17.2 ***Integrated Theory of Influence***

Source: From "Reconceptualizing Evaluation Use: An Integrated Theory of Influence" (p. 8) by K.E. Kirkhart, 2000. In V.J. Caracelli & H. Preskill (Eds.), *The Expanding Scope of Evaluation Use*, New Directions for Evaluation, No. 88. San Francisco: Jossey-Bass. Reprinted with permission.

traditionally, evaluators have focused on use, whether instrumental or conceptual, resulting from the findings of the evaluation study. But Kirkhart's model includes a dimension for the influence of the evaluation process.

The dimension of *intention* alerts us to consider unintended ways in which evaluations are used. In Chapters 11 through 14, we recommend and discuss ways to plan an evaluation by identifying the purposes of the study, the intended audiences, and the potential uses they might make of the results. All evaluations begin this way, with intentions or plans for influence.[1] Such plans can help increase the influence of evaluation by recognizing the needs of different stakeholders. But, Kirkhart notes: "Attention to the unintended influence of evaluation acknowledges both the power of ripple effects and our inability to anticipate all ramifications of our work" (2000, p. 12). She believes that unintended influences may actually exceed intended influences in many settings.

[1]Exceptions would include evaluations that are intended to meet the minimal requirements of a mandated evaluation. In some such cases, the purpose is ill-defined accountability achieved by completing the required forms. When we have worked with such evaluations, we often are able to expand or tweak the evaluation to have some potential use for other stakeholders—often program managers or deliverers.

Finally, her dimension of *time* reminds us that immediate or end-of-cycle influences are not the only occasions when results might be used. While time is presented in the model as three discrete categories, Kirkhart observes that this dimension is actually a continuum. The recognition of long-term influence reminds us that our evaluations may have effects—intended or unintended, due to the findings of the study or the effect of participating in it—long after the evaluator has left the scene. Examples of such influence abound: the principal who makes use of an element of an evaluation process used five years earlier at her previous school to bring parents and teachers together for planning, the social worker who recalls an effect of an early intervention program on toddlers from an evaluation she participated in at a previous agency and uses it with new clients, the manager who adopts a routine use of focus groups and interviews with clients for feedback based on experience in a prior evaluation. Each of these examples illustrates the long-term influence of evaluation.

Prior conceptions of use would have constituted only one or two cubes in the integrated model of influence: intended use occurring immediately or at the end of a cycle. Kirkhart's integrated theory of influence extends earlier thinking, requiring evaluators to consider the wide array of influences that evaluation can have.

Henry and Mark (2003) have proposed a different model, or framework, that is intended to guide research and practice on the effects of evaluation. Like Kirkhart, Henry and Mark feel we should move beyond the term "use" to examine the *effects* of evaluation. Ironically, they believe that the emphasis on immediate use may have prevented or hindered some long-term uses and that it has definitely hindered our examination of other types of uses. Their model postulates three levels, or categories, of influence and, within each, they list some types of influence or change that can occur at that level. (See Figure 17.3.) These levels and types of change are drawn from research findings about change in the fields of psychology, political science, organizational behavior, and others. Their model alerts us to consider evaluation's impact not only on individuals and their attitudes and beliefs, but also on interactions and, eventually, on collective actions in groups, organizations, corporations, governmental units, and the like.

In the aspects discussed so far, Henry and Mark's model is analogous to Kirkhart's integrated theory of influence in that it simply identifies other categories of influence. This, in itself, is useful. However, these authors' real emphasis is on studying *pathways* that lead to use. They argue that evaluation theory and literature have articulated different types of use, but they have not examined how different types of uses are achieved. To make their point, they provide examples of two potential pathways to illustrate evaluation influence. One begins with evaluation results, the other with an influence of process.

Path One: Evaluation Findings → Minority Group Influence → Agenda Setting → Opinion Elaboration → Policy Change

In this hypothetical pathway, evaluation findings prompt a group whose opinion differs from that of the majority to take those findings and make others—the

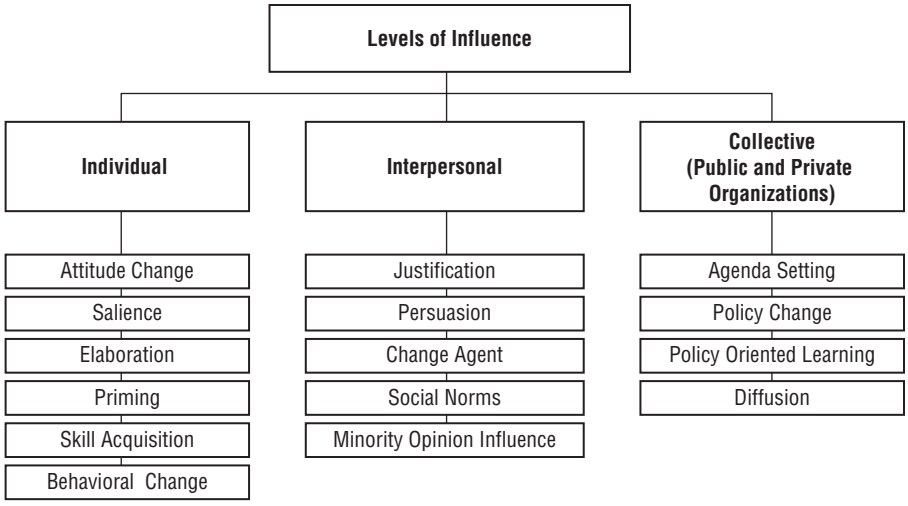


FIGURE 17.3 ***Mechanisms Through Which Evaluation Produces Influences***

Source: From "Beyond Use: Understanding Evaluation's Influence on Attitudes and Actions" (p. 298) by G.T. Henry and M.M. Mark, 2003, *American Journal of Evaluation, 24*(4), 293–314.

public, the media, or an organization—aware of the issue.[2] As the public becomes more aware of the issue, they consider it further and develop, refine, and elaborate their opinions of the issue. The evaluation findings play a role in that elaboration. Finally, the resulting, newly defined opinion held by the public prompts a policy change.

Path Two: Evaluation Process → Attitude Change → Social Norms → Behavior Change

In this pathway, participation in an evaluation process brings about an attitude change in an individual. For example, through participating in an evaluation, an administrative leader—a principal or a manager of a nonprofit organization—may become more sensitive to the need to involve other stakeholders in decisions. She may discuss this with other managers in her organization and gradually bring about a change in social norms about participation. Ultimately, the change in social norms may result in a behavior change at the organizational level to include more participation in decisions.

These hypothetical pathways, and others, are useful because they help us to think of the *mediating steps* that are necessary to bring about change as a result of evaluation. How can practicing evaluators use these ideas? They might consider what the desired end goals are and then brainstorm potential pathways to that goal. By identifying the first step in the pathway, evaluators may be able to more appropriately design the evaluation and reporting process to encourage the type

[2]"Minority group influence" does not refer to a demographic minority, but to a group that holds an opinion that differs from the majority opinion.

of change identified in the first step (as, for example, attitude change or minority group activation in the two pathways).

Steps to Take to Influence Evaluation Use

Over the years, we have learned some things about use. We have expanded and clarified some different ways in which evaluation can influence policy, programs, individuals, or organizations. We have developed some models that can guide us in thinking about influence or pathways to bring about specific types of influence or use. Research on knowledge utilization as well as evaluation use has informed us. Following, then, are a set of recommendations we have drawn from these different sources to advise evaluators on maximizing use:

- Know and work to understand not only the program and the organization in which it resides, but also the policy and decision-making context (Leviton, 2003).

- Consider whether the context lends itself to using participatory or collaborative approaches to increase the potential users' understanding of the evaluation so that either the findings or the process will have an impact on those users (Cousins, 2003; Cousins & Shula, 2006).

- Consider finding one or more key users with the power and the interest to make use of the study's findings, and involve and inform them often on the evaluation, making them partners in your work (Patton, 1997c, 2008a).

- Learn about the values, beliefs, experiences, and knowledge of key users regarding the subjects of the evaluation, and when your results do not correspond to their truth tests consider how to overcome that barrier. For example, incorporate other stakeholders or involve them more in the data collection, analysis, and interpretation of results. Make sure the report passes the utility test as well—that it appears useful.

- Basic as it may sound, increase the frequency and meaningfulness of contacts with key users and other stakeholders. Much of getting evaluation used involves talking with people, as we have emphasized in the earlier section of this chapter on reporting. Don't always be the expert, only giving information. Get information from them. Share and exchange to learn more and to equalize status (Huberman, 1994).

- Learn about the networks within and outside of the organization. Social networks are a major area of research today. Using interpersonal networks and social processing (having others process the information and interpret it) can increase knowledge of the study, and both process and findings use. (See *Evaluation in Action* interview with Jean King for an example of social processing to interpret results. Other examples include Cousins and Leithwood [1993] and Mycio-Mommers [2002]).

- Consider Kirkhart's model of influence and the pathways described by Henry and Mark to identify particular ways in which your team hopes to bring about influence.

In the beginning of this chapter, we presented recommendations for reporting results to encourage use. Like Cook (1997), we believe that direct, instrumental use does occur and that evaluators should encourage such use. We also believe that the expanded definitions of evaluation use or influence can help us to consider the myriad ways our results can influence individuals and institutions. Active consideration of these influences helps us to conduct evaluations and to disseminate results in such a way that evaluation can achieve its purposes of helping others make judgments of merit and worth that ultimately lead to the bettering of society.

Reporting and Influence

This chapter has focused on how to present final results using both written and oral formats to maximize stakeholders' understanding of the results. We encourage you to follow our guidelines. There is nothing wrong with working to stimulate understanding and appropriate actions at the conclusion of the study. However, remember that such actions are but one form of evaluation's influence. As Kirkhart believes, much of evaluation's influence is unintended and may be based on the process rather than the results or conceptual uses of the findings, and it might not occur until long after the study is completed. The Joint Committee *Standards* require utility, meaning the results must have the potential to be used. They do not require immediate use because they recognize that many issues are out of the control of evaluators. Henry (2000), in fact, argues that focusing on immediate use may deter long-term, more important actions (e.g., policy changes) that take time to accomplish. We support the broader interpretation of evaluation's use by considering its influence on many different individuals and institutions in many different ways.

Major Concepts and Theories

1. The reporting of evaluation results is an ongoing process that should typically not be delayed until the end of the evaluation. Evaluators should take advantage of opportunities to report and discuss findings with key intended users and other stakeholders as the evaluation progresses to gain acceptance of the findings, learn of the intended users' reactions, and, ultimately, achieve use.

2. Evaluation reports can serve many different purposes. Consider the purpose for each audience and the means and method of reporting that will best achieve that purpose.

3. Reporting should be adapted to the characteristics and expectations of each audience. An audience analysis can be performed to determine their information needs, perceptions, and values regarding the program and how they receive, respond to, and use the evaluation information. This will help the evaluator to select the best forum

and the most appropriate styles with which to deliver information and engage in dialogue.

4. Involving audiences in decisions about timing, mode, and types of information presented can increase their receptivity to the results.

5. Findings from an evaluation can be presented in many different ways, including short, written communications distributed at meetings; e-mail, web sites, or blogs; interim and final reports; brochures or newsletters; oral presentations supplemented with visuals; posters; working sessions; and personal discussions.

6. Other factors to consider in reporting evaluation findings include accuracy and balance, timing, communication styles and tone, and ways to involve the audience and stimulate dialogue.

7. Final reports include an executive summary, an introduction, a description of the evaluation object and the questions to be answered; a brief discussion of the evaluation plan and methods; a presentation of results; and a discussion of conclusions and recommendations. Detailed technical information can be presented in an appendix.

8. Evaluation use may be better conceptualized as influence. Some evaluations result in a direct, immediate, and intended use of results, but other uses, more broadly categorized as influence, have also emerged. These include the influences of the evaluation process, unintended influences, and influences that occur long after the evaluation has been completed.

Discussion Questions

1. When might an evaluator not want to share interim results? Why?

2. Discuss the advantages and potential disadvantages of using e-mail as your medium for communicating results of an evaluation. What audiences might you involve in this way? How might you stimulate dialogue and interaction? What about learning? How could you encourage feedback?

3. Contrast the effectiveness of an internal evaluator and an external evaluator in reporting results to different stakeholders. What advantages and disadvantages does each role have?

4. Many of the American Evaluation Association's Guiding Principles concern the reporting and use of evaluation. (See Appendix A, in particular A3, C5, D4, and E3, 4, and 5.) (a) Discuss the implications of these principles for reporting. (b) Consider the last evaluation you were involved in. Did the reporting meet the intent of these principles?

5. How do you best receive information? Do you prefer written or oral reports? Shorter or longer briefings or reports? Quantitative or qualitative data? Formal or informal styles?

6. What type of evaluation use or influence do you think is more important—the influence of the process or of the findings? Why?

Application Exercises

1. Apply the good evaluation report checklist from this chapter and the AEA Guiding Principles to an evaluation study of your choosing. In addition to reading the written materials, consider interviewing one or two key intended users to learn how they received information and how it affected them. Identify the strengths and weaknesses of the reporting process in this evaluation.
2. In class, give a 5 to 10 minute oral report, with visual aids such as PowerPoint slides, on an evaluation. Designate the audience to role-play different stakeholder groups, such as funders, managers, program deliverers, clients, or a citizen group. Design your presentation for this audience. Then, after your presentation, use 5–10 minutes for questions and answers. Finally, get feedback from the audience, who now revert to their role as students. What do they see as the strengths of your presentation for this audience? What would they change?

Relevant Evaluation Standards

We consider the following evaluation standards, which are listed in their entirety in Appendix A, to be relevant to this chapter's content:

U1—Evaluator Credibility
U2—Attention to Stakeholders
U3—Negotiated Purposes
U4—Explicit Values
U5—Relevant Information
U6—Meaningful Processes and Products
U7—Timely and Appropriate Communicating and Reporting
U8—Concern for Consequences and Influence
F3—Contextual Viability
P1—Responsive and Inclusive Orientation
P4—Clarity and Fairness
P5—Transparency and Disclosure
A1—Justified Conclusions and Decisions
A7—Explicit Evaluation Reasoning
A8—Communication and Reporting

Case Studies

For this chapter, we recommend three interviews that illustrate different aspects of reporting. *Evaluation in Action*, Chapters 3 (Jennifer Greene), 9 (Stewart Donaldson), and 12 (Bledsoe).

In Chapter 3, Greene works with clients with whom she has worked in the past. She discusses different ways she communicates with them during the project and at the conclusion. Greene and Fitzpatrick also discuss her inclusion of a statement regarding environmental values in the final report, an important part of Greene's approach to evaluation. The journal source is Fitzpatrick, J. L., & Greene, J. C. (2001). Evaluation of the Natural Resources Leadership Program: A dialogue with Jennifer Greene. *American Journal of Evaluation*, *22*(1), 81–96.

In Chapter 9, Donaldson describes his work with several different clients in this evaluation, though his primary focus is discussing one particular project. He discusses ongoing communication with the stakeholders and the funder, including changing roles from formative to summative as the project progresses and using 360 degree feedback—having stakeholders provide feedback to the evaluators on their performance. The journal source is Fitzpatrick, J. L., & Donaldson, S. I. (2002). Evaluation of the Work and Health Initiative: A dialogue with Stewart Donaldson. *American Journal of Evaluation*, *23*, 347–365.

In Chapter 12, Bledsoe brings about significant instrumental and conceptual use. She describes the influence of the findings and the process—but consider how she brings about users' openness to change. The journal source is Fitzpatrick, J. L., & Bledsoe, K. (2007). Evaluation of the Fun with Books Program: A dialogue with Katrina Bledsoe. *American Journal of Evaluation*, *28*, 522–535.

Suggested Readings

Henry, G. T. (Ed.). (1997). *Creating effective graphs: Solutions for a variety of evaluation data.* New Directions for Evaluation, No. 73. San Francisco: Jossey-Bass.

Torres, R. T., Preskill, H., & Piontek, M. (2005). *Evaluation strategies for communicating and reporting: Enhancing learning in organizations* (2nd ed.). Newbury Park: Sage.

18

The Future of Evaluation

Orienting Questions

1. How are future program evaluations likely to be different from current evaluations in
   - the way in which political considerations are handled?
   - the approaches that will be used?
   - the involvement of stakeholders?
   - who conducts them?
2. How is evaluation like some other activities in organizations?
3. How is evaluation viewed differently in other countries?

We have reached the last chapter of this book, but we have only begun to share what is known about program evaluation. The references we have made to other writings reflect only a fraction of the existing literature in this growing field. In choosing to focus attention on (1) alternative approaches to program evaluation, and (2) practical guidelines for planning, conducting, reporting, and using evaluation studies, we have tried to emphasize what we believe is most important to include in any single volume that aspires to give a broad overview of such a complex and multifaceted field. We hope we have selected well, but we encourage students and evaluation practitioners to go beyond this text to explore the richness and depth of other evaluation literature. In this final chapter, we share our perceptions and those of a few of our colleagues about evaluation's future.

The Future of Evaluation

Hindsight is inevitably better than foresight, and ours is no exception. Yet present circumstances permit us to hazard a few predictions that we believe will hold true for program evaluation in the next few decades. History will determine whether

or not our predictions are accurate enough to add prophecy to the repertoire of techniques useful to evaluators.

We believe that evaluation will continue to spread rapidly around the globe, until there are few countries, territories, provinces, states, and locales in which program evaluations are not at least an occasional occurrence. As we have noted, the spreading interest in program evaluation has been evident for some years in the development of evaluation associations and activities around the world. We also believe that evaluation will become an increasingly useful force in the following ways:

- Improving programs, thus improving the lot of those intended to benefit from those programs
- Improving policy making by governing boards, legislators, and congressional and parliamentary bodies
- Improving organizational learning and decision making
- Improving societies through improving the social conditions programs address
- Improving even itself

If these predictions seem overly optimistic, it may underscore our earlier point that evaluators may not always be completely unbiased. Yet these forecasts do not strike us as unrealistic or overdrawn; we are willing to submit them to the test of time.

Now let us move to more specific predictions concerning the profession of evaluation and its practice (Worthen, 2001).

Predictions Concerning the Profession of Evaluation

1. Evaluation will become an increasingly useful force in our society. As noted, evaluation will have increasing impacts on programs, on organizations, and on society. Many of the movements we have discussed in this text—performance monitoring, organizational learning, and others—illustrate the increasing interest in and impact of evaluation in different sectors. Evaluative means of thinking will improve ways of planning and delivering programs and policies to achieve their intended effects and, more broadly, improve society.

2. Evaluation will increase in the United States and in other developed countries as the pressure for accountability weighs heavily on governments and nonprofit organizations that deliver vital services. The emphasis on accountability and data-based decision making has increased dramatically in the first decade of the twenty-first century. Also, virtually every trend points to more, not less, evaluation in the public, private, and nonprofit sectors in the future. In some organizations, the focus is on documenting outcomes in response to external political pressures. In other organizations, evaluation is being used for organizational growth and development, which should, ultimately, improve the achievement of those outcomes. In each context, however, evaluation is in demand.

3. Evaluation will continue to spread rapidly around the globe until there are few countries where program evaluations are not at least an occasional occurrence. In addition, the number of national and multinational professional societies for evaluation will burgeon.

4. The opportunity for careers in evaluation will continue to increase as the demand for evaluation skills grow. As LaVelle and Donaldson write in 2010, "Evaluation practice has grown in leaps and bounds in recent years" (2010, p. 9). The growth in evaluation in the United States is demonstrated by the dramatic growth in membership of the American Evaluation Association in the last decade (Mark, 2007).

5. Graduate programs in evaluation will increase with the growing demand. Lavelle and Donaldson (2010) found that the number of graduate training programs in evaluation increased to 48 in 2008, a dramatic increase since 2006 when there were only 27 such programs following several years of decline. Increases were particularly notable in schools of education, but programs were also found in public policy, psychology, criminal justice, and sociology. More than half of the 48 programs, however, were small ones, offering only two to three evaluation-specific courses. As the profession grows, and the market demands more professionals trained specifically in evaluation, we hope the incidence and depth of such programs grow as well. A world with growing evaluation demands requires evaluations that are conducted by people trained in the many options that evaluators can pursue.

6. Many of those conducting evaluations will need more specific evaluation training. Graduate programs cannot keep up with the demand. Further, since evaluation is a relatively new profession, many are not aware of the profession and the specific approaches and methodologies professional evaluators use. Many of those conducting evaluations within organizations, and as external consultants, continue to be trained in social science methodologies, or in organizational management, but have not received in-depth evaluation training. As educators, public and nonprofit administrators, corporate officers, and those in a variety of other roles are asked to assume some responsibility for carrying out evaluation studies along with their other professional duties, the need for in-service education in evaluation will grow (Datta, 2006).

7. Internal evaluation will, despite its risks, become more important because of its benefits. Internal evaluators know the organizational environment. They can provide an important ongoing influence to encourage organizational learning and to use evaluation skills across the organization in many different endeavors, from using new information technology to human resource management and traditional evaluation of programs. We predict there will be increased cooperation between internal and external evaluators in many evaluations.

8. Professional associations will continue to grow and to branch into new areas to expand the public presence of evaluation. In 2010, membership in the American Evaluation Association (AEA) should reach 6,000 (Kistler, 2010).

The AEA and other societies of practicing evaluators and/or evaluation theoreticians will continue to contribute to evaluation's maturation. Like other professional associations, the AEA has begun taking public positions on issues of relevance to evaluation and has recently begun to direct its attention to evaluation policy and working to influence such policies. (See Trochim, Mark, and Cooksy [2009]) The Canadian Evaluation Society has developed a program to recognize professional evaluators through a credentialing process that recognizes those who apply and meet the criteria as Credentialed Evaluators. As in other professions, this process is intended to help clients and stakeholders to distinguish professional evaluators from those with less direct training or experience. Through these efforts and continuing education regarding evaluation practice and standards, associations will work to educate stakeholders about the potential that evaluation offers and the risks entailed in using it inappropriately.

9. Evaluation literature will increase in both quantity and quality, but relatively little of it will be based on research into the process of evaluation itself. Current funding agencies do not seem interested in supporting research concerning the evaluation process. Governments are investing many resources in accountability, performance monitoring, and evaluation to determine if programs work and how they work. However, our approaches to these subjects continue to rely on reasoning and intuition rather than solid evidence about how evaluators can best work with stakeholders, what forms of participation lead to what types of impacts and use, and so forth. Therefore, the empirical knowledge base in evaluation will increase very slowly. As evaluation expands, there is a critical need for more research on what occurs, what works, and what doesn't in evaluation practice, participation, and use.

Predictions Concerning the Practice of Evaluation

As the profession grows and expands, practice will change even more dramatically.

1. Approaches to evaluation will become more eclectic and adaptive to contextual circumstances. Program evaluation will continue to be pluralistic, but will move toward greater pragmatism as evaluators work to provide valid and appropriate findings and conclusions and to improve programs, policies, and decision making in many different settings. Single-method evaluations will be viewed by professional evaluators, if not by the public and some elected officials, as simplistic and inadequate for evaluation of complex programs or those serving diverse populations. Triangulation, cross-validation, and iterative, expansive designs will be used more routinely to allow the complementarity of qualitative and quantitative approaches to enrich evaluation work. The usefulness of the different approaches will lie less in having any one of them serve as a model to be followed slavishly but rather, as House (1994a) has suggested, as collectively

comprising the "grammar of evaluation" that evaluators must understand and be skilled in using:

> [One] might see the evaluation models as something like model sentences in a grammar of evaluation. . . . As one progresses, . . . one does not need to think about the models consciously, except to correct particular errors or study the grammar itself.
>
> Similarly, . . . experienced evaluators can construct evaluation designs which do not depend explicitly on particular models. Actual evaluation designs can be combinations of elements from different models, . . . just as speakers can produce novel grammatical sentences once they have learned the basic grammar of a language. (pp. 241–242)

2. Evaluation will be mainstreamed in organizations. Some of its methods, such as logic models and program theories, are already being used by program managers and staff to develop programs. As process learning from evaluation and organizational learning increase, evaluative ways of thinking in organizations will expand. Evaluation won't always be called *evaluation*, but its influence will be felt through creating a culture of learning and using information and data to make decisions (Mark, 2007).

3. Evaluation will expand to evaluate programs in new arenas. In the United States and Canada, evaluators have come primarily from education and psychology and have evaluated programs in those areas. But, the role of evaluation in housing, social welfare, environmental programs, city planning, transportation, health, criminal justice, biotechnology, recreation, and environmental programs continues to expand. Working in these new areas will prompt evaluators to expand their repertoire of approaches and methods to adapt to these new contexts, new political dynamics, and new issues to explore and investigate. In Europe, evaluators typically come from the disciplines of political science, economics, and public administration and, as a result, focus their efforts on different types of programs using somewhat different methods. Our growing awareness of these differences makes us realize that evaluators in each country can learn from practice in other countries.

4. Evaluators will become more aware of and involved in the work of planners, policy analysts, and organizational developers. Evaluation activities overlap with the work of policy analysts, planners, and organizational developers. Evaluators have approaches and qualitative methods that could contribute to the work of policy analysts. They have economic and analytic methods that could add to the repertoire of evaluators. Similarly, city planners and program planners collect information in ways that are similar to an evaluator conducting a needs assessment. Evaluators' skills in developing logic models and program theory can help in program planning. We predict that more communication will occur across these fields, with professionals sharing approaches and methods. As noted previously, the work they do may not be called *evaluation*, but evaluation professionals will be bringing their evaluative skills to the tasks.

5. Evaluation (and evaluators) will become more politically sophisticated, recognizing that our goal is to encourage policymakers and program managers to use evaluative information to make decisions and to educate voters and the

public. As we move to more eclectic and adaptive evaluation practices and avoid our own infighting, we can be more successful in this venture. Greene and Henry advise us to recognize that we do not want evaluation's disputes to "become a license for actions based entirely on ideology or the force of unconstrained rhetoric. We should unite in our commitment . . . to reclaim the conversation about the contributions of social science to social policies and programs" (2005, p. 350). Ideology and rhetoric are part of the political system, just as is public opinion. Information, conclusions, and judgments provided by evaluation studies become one piece of this mix of input that policymakers receive. For our activities to rise to the fore, evaluators must balance the perceived objectivity of their roles against entering the political fray to call attention to information that can improve programs and policy. Today, evaluation studies compete with information supplied by partisan, overtly political think tanks. Citizens, and some policymakers, may not know the difference. Evaluators must be skilled at helping stakeholders recognize the strengths of our work.

6. Attention to ethical issues will increase as evaluators become more involved in political issues. The present evaluation *Standards* and Guiding Principles—and their descendants—provide a means for maintaining the credibility of evaluation and educating others about codes and ethics in an increasingly politicized environment. Professional accountants have strengthened their ethical codes and training of practitioners in the face of public disillusionment concerning the "independent" role of accountants in reviewing the financial practices of corporations (Fitzpatrick, 1999). Evaluators can avoid the debacle that the Arthur Andersen Accounting firm faced in its auditing of Enron, and those that Standard and Poors and other bond rating firms endured as their high-rated bonds fell in the economic crisis of 2009, by strengthening the ethical education of current and future evaluators.

7. Electronic and other technological advances will alter the way evaluators collect information, draw conclusions, and report findings, enabling broader stakeholder participation and access to evaluation reports and their findings. Today, data are routinely collected through Internet surveys and interviews. Results can be shared and discussed online with members of the evaluation team and advisory councils to consider interpretations. Databases can be shared so members can explore the data for different interpretations. Interim and final reports are routinely posted on organizational web sites with links to videos and audio depicting the program or demonstrating results. Readers are encouraged to post comments. Such uses, and others as yet unanticipated, will increase as technological capacity and literacy increase.

8. Efforts will increase to democratize evaluation as part of the movement in many countries for more citizen input. Across the United States, deliberative democracy methods are being used, where local citizens serve with elected officials and others to learn about policy issues and make recommendations. Participatory evaluation is part of that movement. We will continue to democratize evaluation by involving many different stakeholders and educating them on evaluative ways of thinking and purposes.

9. The performance measurement movement will grow in response to persistent demands for accountability. Performance measurement, in some form or fashion, is now mandated in most local, state, and federal government agencies and in nonprofit organizations led by initiatives from United Way and the World Bank. Expectations from the public and from policymakers who mandated performance measurement are high. Yet most managers lack the expertise to collect meaningful measures of outcomes. Newcomer (2001) notes that professional evaluators will play an important role in making this process more than simply a reporting exercise. Evaluators can help build program theory to link outcomes to program activities and, hence, make the outcome information useful for formative purposes. Further, evaluators' methodological expertise will be necessary to measure outcomes.

Performance measurement, however, also presents potential hazards for the evaluation field. Just as states' testing of students on educational standards has grossly simplified learning goals and focused educational evaluation activities on just this issue, so, too, can performance measurement simplify and narrow evaluation activities. Many policymakers and managers underestimate the challenge of measuring program outcomes and, because of the mandated nature of performance measurement, tend to see performance measurement as all that evaluation does. Evaluators need to be active in this area to bring their expertise to bear.

A Vision for Evaluation

In addition to the predictions we have for the profession and practice of evaluation, we also have some visions, or goals, for it. These differ from predictions in that the evidence is not so clear as to whether these visions will be realized. Nevertheless, we would be remiss if we ended this book without describing that vision. It includes the following:

1. A global valuing of evaluation that cuts across boundaries of professional fields, job categories, sectors of society, geopolitical lines, cultures—that is, formal disciplined evaluation as a pervasive value. How will we bring about this valuing? By making others aware of evaluation and its importance. By helping those who are mandated to do evaluations to see its worth even when it is not mandated. By instilling evaluation institutions, policies and procedures, and evaluative ways of thinking in organizations (Sanders, 2001).

2. Continuing a constructive use of multiple methods and eclectic approaches to achieve the many different purposes of evaluation. The debates over qualitative and quantitative methods have subsided and many have moved on to the practical issues of applying their now increased methodological tools in a variety of settings. To avoid future divisive debates we should recognize the plurality of evaluation purposes, questions, and settings. An evaluator working with the U.S. government on performance monitoring issues is facing different methodological and political challenges than the evaluator designing a special, formative

study for a nonprofit agency on its work with a new group of clients. Rather than debate the different choices these evaluators make, we should study their choices and learn more about what approaches work best in different settings. As evaluators, we should know not to judge decisions made in other evaluations without sufficient information. We need to work harder to defer that judgment and explore and collect information on those choices. Let's develop thick descriptions about evaluations!

3. Increase the use of metaevaluation to improve evaluation practice. One type of publication that is regrettably rare in the evaluation journals is critiques of prior evaluation reports, that is, metaevaluations. Despite the acceptance and availability of the Joint Committee's *Standards*, few evaluations appear to be subjected to any closer scrutiny today than before metaevaluation standards were developed. To learn from our own work, we must be open to its review and evaluation. As others learn from our evaluations, evaluators can learn from evaluations of their own work.

Conclusion

We leave the reader with two final thoughts.

First, all our years of experience conducting and studying it convinces us that evaluation, properly conducted, has great potential for improving programs and practices in education, human services, nonprofit organizations—virtually every area of society. Managers, policymakers, and other stakeholders have become aware that some evaluation studies are misused or ignored, with the result that some individuals have argued for decreased emphasis on the evaluative process. But that seems no more sensible than abandoning medical diagnosis because science has not yet succeeded in eliminating all disease.

The second thought we wish to leave with readers is this: Despite great strides, it is increasingly apparent how little we really do know about evaluation, compared with what we need to know. It is our earnest hope that this book has added to that knowledge and has helped to illuminate the thousand points of darkness that still constitute current processes for creating and implementing the policies and programs intended to improve the lot of humankind.

Suggested Readings

Datta, L. (2006). The practice of evaluation: Challenges and new directions. In J. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation*. Thousand Oaks, CA: Sage.

Mark, M. M. (2007). AEA and evaluation: 2006 (and beyond). In S. Mathison (Ed.), *Enduring issues in evaluation: The 20th anniversary of the collaboration between NDE and AEA*. New Directions for Evaluation, No. 114. San Francisco: Jossey-Bass.

Smith, M. F. (2001). Evaluation: Preview of the future #2. *American Journal of Evaluation*, *22*, 281–300.

See the entire issues of *American Journal of Evaluation* (2002), 22, and *Evaluation Practice* (1994), 15. Each of these issues focuses on reflections concerning evaluation theories, practice, and status as a profession and predictions concerning the future.

This page intentionally left blank

Appendix A

The Program Evaluation Standards and *Guiding Principles for Evaluators*

The Program Evaluation Standards

Utility

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 re-discover, re-interpret 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.

Feasibility

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 effectively and efficiently.

Propriety

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.

Accuracy

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.

Evaluation Accountability

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.

Source: From *The Program Evaluation Standards,* 3rd ed., by D. B. Yarbrough, L. M. Shulha, R. K. Hopson, & F. A. Caruthers, 2011. [Approved by the Joint Committee on Standards for Educational Evaluation, an ANSI-Accredited Standards Developer.] Thousand Oaks, CA: Sage. Used with permission.

American Evaluation Association's Guiding Principles for Evaluators

A. Systematic Inquiry: Evaluators conduct systematic, data-based inquiries.

1. To ensure the accuracy and credibility of the evaluative information they produce, evaluators should adhere to the highest technical standards appropriate to the methods they use.
2. Evaluators should explore with the client the shortcomings and strengths both of the various evaluation questions and the various approaches that might be used for answering those questions.
3. Evaluators should communicate their methods and approaches accurately and in sufficient detail to allow others to understand, interpret and critique their work. They should make clear the limitations of an evaluation and its results. Evaluators should discuss in a contextually appropriate way those values, assumptions, theories, methods, results, and analyses significantly affecting the interpretation of the evaluative findings. These statements apply to all aspects of the evaluation, from its initial conceptualization to the eventual use of findings.

B. Competence: Evaluators provide competent performance to stakeholders.

1. Evaluators should possess (or ensure that the evaluation team possesses) the education, abilities, skills and experience appropriate to undertake the tasks proposed in the evaluation.
2. To ensure recognition, accurate interpretation and respect for diversity, evaluators should ensure that the members of the evaluation team collectively demonstrate cultural competence. Cultural competence would be reflected in evaluators seeking awareness of their own culturally-based assumptions, their understanding of the worldviews of culturally-different participants and stakeholders in the evaluation, and the use of appropriate evaluation strategies and skills in working with culturally different groups. Diversity may be in terms of race, ethnicity, gender, religion, socio-economics, or other factors pertinent to the evaluation context.
3. Evaluators should practice within the limits of their professional training and competence, and should decline to conduct evaluations that fall substantially

outside those limits. When declining the commission or request is not feasible or appropriate, evaluators should make clear any significant limitations on the evaluation that might result. Evaluators should make every effort to gain the competence directly or through the assistance of others who possess the required expertise.

4. Evaluators should continually seek to maintain and improve their competencies, in order to provide the highest level of performance in their evaluations. This continuing professional development might include formal coursework and workshops, self-study, evaluations of one's own practice, and working with other evaluators to learn from their skills and expertise.

C. Integrity/Honesty: Evaluators display honesty and integrity in their own behavior, and attempt to ensure the honesty and integrity of the entire evaluation process.

1. Evaluators should negotiate honestly with clients and relevant stakeholders concerning the costs, tasks to be undertaken, limitations of methodology, scope of results likely to be obtained, and uses of data resulting from a specific evaluation. It is primarily the evaluator's responsibility to initiate discussion and clarification of these matters, not the client's.
2. Before accepting an evaluation assignment, evaluators should disclose any roles or relationships they have that might pose a conflict of interest (or appearance of a conflict) with their role as an evaluator. If they proceed with the evaluation, the conflict(s) should be clearly articulated in reports of the evaluation results.
3. Evaluators should record all changes made in the originally negotiated project plans, and the reasons why the changes were made. If those changes would significantly affect the scope and likely results of the evaluation, the evaluator should inform the client and other important stakeholders in a timely fashion (barring good reason to the contrary, before proceeding with further work) of the changes and their likely impact.
4. Evaluators should be explicit about their own, their clients', and other stakeholders' interests and values concerning the conduct and outcomes of an evaluation.
5. Evaluators should not misrepresent their procedures, data or findings. Within reasonable limits, they should attempt to prevent or correct misuse of their work by others.
6. If evaluators determine that certain procedures or activities are likely to produce misleading evaluative information or conclusions, they have the responsibility to communicate their concerns and the reasons for them. If discussions with the client do not resolve these concerns, the evaluator should decline to conduct the evaluation. If declining the assignment is unfeasible or inappropriate, the evaluator should consult colleagues or relevant stakeholders about other proper ways to proceed. (Options might include discussions at a higher level, a dissenting cover letter or appendix, or refusal to sign the final document.)
7. Evaluators should disclose all sources of financial support for an evaluation, and the source of the request for the evaluation.

D. Respect for People: Evaluators respect the security, dignity and self-worth of respondents, program participants, clients, and other evaluation stakeholders.

1. Evaluators should seek a comprehensive understanding of the important contextual elements of the evaluation. Contextual factors that may influence the results of a study include geographic location, timing, political and social climate, economic conditions, and other relevant activities in progress at the same time.
2. Evaluators should abide by current professional ethics, standards, and regulations regarding risks, harms, and burdens that might befall those participating in the evaluation; regarding informed consent for participation in evaluation; and regarding informing participants and clients about the scope and limits of confidentiality.
3. Because justified negative or critical conclusions from an evaluation must be explicitly stated, evaluations sometimes produce results that harm client or stakeholder interests. Under this circumstance, evaluators should seek to maximize the benefits and reduce any unnecessary harms that might occur, provided this will not compromise the integrity of the evaluation findings. Evaluators should carefully judge when the benefits from doing the evaluation or in performing certain evaluation procedures should be foregone because of the risks or harms. To the extent possible, these issues should be anticipated during the negotiation of the evaluation.
4. Knowing that evaluations may negatively affect the interests of some stakeholders, evaluators should conduct the evaluation and communicate its results in a way that clearly respects the stakeholders' dignity and self-worth.
5. Where feasible, evaluators should attempt to foster social equity in evaluation, so that those who give to the evaluation may benefit in return. For example, evaluators should seek to ensure that those who bear the burdens of contributing data and incurring any risks do so willingly, and that they have full knowledge of and opportunity to obtain any benefits of the evaluation. Program participants should be informed that their eligibility to receive services does not hinge on their participation in the evaluation.
6. Evaluators have the responsibility to understand and respect differences among participants, such as differences in their culture, religion, gender, disability, age, sexual orientation and ethnicity, and to account for potential implications of these differences when planning, conducting, analyzing, and reporting evaluations.

E. Responsibilities for General and Public Welfare: Evaluators articulate and take into account the diversity of general and public interests and values that may be related to the evaluation.

1. When planning and reporting evaluations, evaluators should include relevant perspectives and interests of the full range of stakeholders.
2. Evaluators should consider not only the immediate operations and outcomes of whatever is being evaluated, but also its broad assumptions, implications and potential side effects.
3. Freedom of information is essential in a democracy. Evaluators should allow all relevant stakeholders access to evaluative information in forms that respect people and honor promises of confidentiality. Evaluators should actively

disseminate information to stakeholders as resources allow. Communications that are tailored to a given stakeholder should include all results that may bear on interests of that stakeholder and refer to any other tailored communications to other stakeholders. In all cases, evaluators should strive to present results clearly and simply so that clients and other stakeholders can easily understand the evaluation process and results.

4. Evaluators should maintain a balance between client needs and other needs. Evaluators necessarily have a special relationship with the client who funds or requests the evaluation. By virtue of that relationship, evaluators must strive to meet legitimate client needs whenever it is feasible and appropriate to do so. However, that relationship can also place evaluators in difficult dilemmas when client interests conflict with other interests, or when client interests conflict with the obligation of evaluators for systematic inquiry, competence, integrity, and respect for people. In these cases, evaluators should explicitly identify and discuss the conflicts with the client and relevant stakeholders, resolve them when possible, determine whether continued work on the evaluation is advisable if the conflicts cannot be resolved, and make clear any significant limitations on the evaluation that might result if the conflict is not resolved.
5. Evaluators have obligations that encompass the public interest and good. These obligations are especially important when evaluators are supported by publicly-generated funds; but clear threats to the public good should never be ignored in any evaluation. Because the public interest and good are rarely the same as the interests of any particular group (including those of the client or funder), evaluators will usually have to go beyond analysis of particular stakeholder interests and consider the welfare of society as a whole.

Source: From http://www.eval.org/Publications/GuidingPrinciples.asp, American Evaluation Association. Approved originally in 1994, revisions approved by members, 2004. Used with permission.

References

Abma, T. A., & Stake, R. E. (2001). Stake's responsive evaluation: Core ideas and evolution. In J. C. Greene & T. A. Abma (Eds.), *Responsive evaluation*. New Directions for Program Evaluation, No. 92, 7–22. San Francisco: Jossey-Bass.

Affholter, D. P. (1994). Outcome monitoring. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation*. San Francisco: Jossey-Bass.

Agar, M. (2000). Border lessons: Linguistic "rich points" and evaluative understanding. In R. K. Hopson (Ed.), *How and why language matters in evaluation*. New Directions for Evaluation, No. 86, 93–109. San Francisco: Jossey-Bass.

Alkin, M. C. (1969). Evaluation theory development. *Evaluation Comment*, 2, 2–7.

Alkin, M. C. (1991). Evaluation theory development: II. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century. Ninetieth Yearbook of the National Society for the Study of Education, Part II.* Chicago: University of Chicago Press.

Alkin, M. C. (Ed.). (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Alkin, M. C. (2005). Utilization of evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation.* Thousand Oaks, CA: Sage.

Alkin, M. C., & Christie, C. A. (2004). An evaluation theory tree. In M. C. Alkin (Ed.), *Evaluation roots* (pp. 12–65). Thousand Oaks, CA: Sage.

Alkin, M. C., & Ellett, F. (1985). Evaluation models and their development. In T. Husèn & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies.* Oxford: Pergamon.

Alkin, M. C. & House, E. (1992). Evaluation of programs. In M. Alkin (Ed.), *Encyclopedia of educational research.* New York: Macmillan.

Alkin, M. C., Stecher, B. M., & Geiger, F. L. (1982). *Title I evaluation: Utility and factors influencing use*. Northridge, CA: Educational Evaluation Associates.

Altschuld, J. W. (2009). *The needs assessment kit.* Thousand Oaks, CA: Sage.

American Evaluation Association. (1995). Guiding principles for evaluators. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators*. New Directions for Program Evaluation, No. 34, 19–26. San Francisco: Jossey-Bass.

American Evaluation Association. (2002). *Position statement on high stakes testing in Pre K–12 Education*. Retrieved March 1, 2010 from http://www.eval.org/hst3.htm.

American Evaluation Association. (2004). *Guiding principles for evaluators.* Retrieved January 21, 2010 from http://www.eval.org/Publications/GuidingPrinciples.asp.

American Evaluation Association. (2006). Public statement: Educational accountability. Retrieved May 29, 2009 from http://www.eval.org/edac.statement.asp.

American Evaluation Association. (2008). *AEA 2007/2008 Internal Scan Report to the Membership by Goodman Research Group*. Retrieved January 22, 2010 from http://www.eval.org/Scan/aea08.scan.report.pdf.

Amie, M. (1995). The Australasian Evaluation Society. *Evaluation, 1*, 124–125.

Andersen, L. A., & Schwandt, T. A. (2001). *Mainstreaming evaluation in society: Understanding the moral/political orientation of evaluation practice.* Paper presented at the annual meeting of the American Evaluation Association, St. Louis, MO.

Arens, S. A., & Schwandt, T. A. (2000). Review of values in evaluation and social research. *Evaluation and Program Planning, 23*(3), 261–276.

Argyris, C., & Schöen, D. A. (1978). *Organizational learning: A theory of action perspective*. Reading, MA: Addison-Wesley.

Babbie, E., Halley, F. S., Wagner, W. E., III, & Zaino, J. (2010). *Adventures in social research: Data analysis using SPSS 17.0 and 18.0 for Windows* (7th ed.). Thousand Oaks, CA: Sage.

Baker, A., & Bruner, B. (2006). *Evaluation capacity and evaluative thinking in organizations* (Vol. 6). Cambridge MA: Bruner Foundation, Inc.

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1986*). Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Barbour, R. (2008). *Doing focus groups. Qualitative research kit*. Thousand Oaks, CA: Sage.

Barnett, W. S. (1996). *Lives in the balance: Age 27 benefit-cost analysis of the High Scope Perry Preschool Program.* Ypsilanti, MI: High/Scope Press.

Barnette, J. J., & Sanders, J. R. (2003). *The mainstreaming of evaluation.* New Directions for Evaluation, No. 99. San Francisco: Jossey-Bass.

Bell, J. B. (2004). Managing evaluation projects. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (2nd ed.). San Francisco: Jossey-Bass.

Bell, W. (1983). *Contemporary social welfare.* New York: Macmillan.

Belmont Report (1979). Department of Health, Education, and Welfare, The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. http://ohsr.od.nih.gov/guidelines/belmont.html.

Benjamin, L. M., & Greene, J. C. (2009). From program to network: The evaluator's role in today's public problem–solving environment. *American Journal of Evaluation, 30*(3), 296–309.

Benkofske, M. (1994, November). *When the qualitative findings are negative.* Paper presented at the annual meeting of the American Evaluation Association, Boston.

Bernard, H. R. (2000). *Social research methods: Qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Bernard, H. R., & Ryan, G. W. (2010). *Analyzing qualitative data: Systematic approaches.* Thousand Oaks, CA: Sage.

Bickman, L. (Ed.). (1987). *Using program theory in evaluation*. New Directions for Program Evaluation, No. 33. San Francisco: Jossey-Bass.

Bickman, L. (Ed.). (1990). *Advances in program theory*. New Directions for Program Evaluation, No. 47. San Francisco: Jossey-Bass.

Bickman, L. (2002). The death of treatment as usual: An excellent first step on a long road. *Clinical Psychology*. *9*(2), 195–199.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I. Cognitive domain*. New York: David McKay.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook of formative and summative evaluation of student learning.* New York: McGraw-Hill.

Boruch, R. (2007). Encouraging the flight of error: Ethical standards, evidence standards, and randomized trials. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation methodology: Building the evidence base for method choice in government sponsored evaluation.* New Directions for Evaluation, No. 113, 55–74. San Francisco: Jossey-Bass.

Botcheva, L., Shih, J., & Huffman, L. C. (2009). Emphasizing cultural competence in evaluation: A process-oriented approach. *American Journal of Evaluation*, *30*(2), 176–188.

Brandon, P. R. (1998). Stakeholder participation for the purpose of helping ensure evaluation validity: Bridging the gap between collaborative and non-collaborative evaluations. *American Journal of Evaluation, 19*, 325–337.

Brandon, P. R. (2005). Using test standard-setting methods in educational program evaluation: Assessing the issue of how good is good enough. *Journal of Multidisciplinary Evaluation*, *3*, 1–29.

Brandon, P. R., Taum, A. K. H., Young, D. B., Pottenger, F. M., III, & Speitel, T. W. (2008). The complexity of measuring the quality of program implementation with observations: The case of middle school inquiry-based science. *American Journal of Evaluation, 29*(2), 235–250.

Braverman, M. T. (1996). Sources of survey error: Implications for evaluation studies. In M. T. Braverman & J. K. Slater (Eds.), *Advances in survey research*. New Directions for Evaluation, No. 70, 17–28. San Francisco: Jossey-Bass.

Braverman, M. T., & Slater, J. K. (1996). *Advances in survey research*. New Directions for Evaluation, No. 70. San Francisco: Jossey-Bass.

Brinkerhoff, R. O. (2003). *The success case method: How to quickly find out what's working and what's not.* San Francisco: Berrett Kohler.

Brinkerhoff, R. O., Brethower, D. M., Hluchyj, T., & Nowakowski, J. R. (1983). *Program evaluation: A practitioner's guide for trainers and educators.* Boston: Kluwer-Nijhoff.

Brittingham, B. (2009). Accreditation in the United States: How did we get to where we are? In P. M. O'Brian (Ed.), *Accreditation: Assuring and enhancing quality.* New Directions in Higher Education, No. 145, 7–27. San Francisco: Jossey-Bass.

Buchanan, G. N., & Wholey, J. S. (1972). Federal level evaluation. *Evaluation, 1*, 17–22.

Buckley, J., & Schneider, M. (2006). Are charter school parents more satisfied with schools? Evidence from Washington, DC. *Peabody Journal of Education, 81*(1), 57–78.

Calhoun, E. F. (1994). *How to use action research in the self-renewing school.* American Society for Curriculum Development.

Calhoun, E. F. (2002). Action research for school improvement. *Educational Leadership, 59*(6), 18–25.

Campbell, D. (1969a). Ethnocentrism of disciplines and the fish-scale model of omniscience. In M. Sheriff & C. Sherif (Eds.), *Inter-disciplinary relationships in the social sciences.* Chicago: Aldine.

Campbell, D. T. (1969b). Reforms as experiments. *American Psychologist, 24,* 409–429.

Campbell, D. T. (1984). Can we be scientific in applied social science? In R. F. Conner, D. G. Altman, & C. Jackson (Eds.), *Evaluation studies review annual* (Vol. 9). Beverly Hills, CA: Sage.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Canadian Evaluation Society. (1992). Standards for program evaluation in Canada: A discussion paper. *Canadian Journal of Program Evaluation, 7,* 157–170.

Caracelli, V. J., & Greene, J. C. (1997). Crafting mixed-method evaluation designs. In J. C. Greene & V. J. Caracelli (Eds.), *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms.* New Directions for Program Evaluation, No. 74. San Francisco: Jossey-Bass.

Carman, J. G., Fredericks, K. A., & Introcaso, D. (2008). Government and accountability: Paving the way for nonprofits and evaluation. In J. G. Carman & K. A. Fredericks (Eds.), *Nonprofits and evaluation.* New Directions for Evaluation, No. 119, 5–12. San Francisco: Jossey-Bass.

Caro, F. G. (Ed.). (1971). *Readings in evaluation research.* New York: Russell Sage.

Carr, W., & Kemmis, S. (1992). *Becoming critical: Education, knowledge, and action research.* London: Falmer.

Chelimsky, E. (1994). Evaluation: Where are we? *Evaluation Practice, 15,* 339–345.

Chelimsky, E. (1997). The coming transformations in evaluation. In E. Chelimsky & W. R. Shadish (Eds), *Evaluation for the 21st Century: A handbook.* Thousand Oaks, CA: Sage.

Chelimsky, E. (1998). The role of experience in formulating theories of evaluation practice. *American Journal of Evaluation, 20,* 35–56.

Chelimsky, E. (2001). What evaluation could do to support foundations: A framework with nine component parts. *American Journal of Evaluation, 22*(1), 13–28.

Chelimsky, E. (2006). The purposes of evaluation in a democratic society. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation.* Thousand Oaks, CA: Sage.

Chelimsky, E. (2007). Factors influencing the choice of methods in federal evaluation practice. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation methodology: Building the evidence base for method choice in government sponsored evaluation.* New Directions for Evaluation, No. 113, 13–34. San Francisco: Jossey-Bass.

Chelimsky, E. (2008). A clash of cultures: Improving the "fit" between evaluative independence and the political requirements of a democratic society. *American Journal of Evaluation, 29,* 400–415.

Chelimsky, E., & Shadish, W. R. (1997). *Evaluation for the 21st century: A handbook.* Thousand Oaks, CA: Sage.

Chen, H. T. (1990). *Theory-driven evaluations.* Newbury Park, CA: Sage.

Chen, H. T. (1994). Current trends and future directions in program evaluation. *Evaluation Practice, 15,* 229–238.

Chen, H. T. (1996). A comprehensive typology for program evaluation. *Evaluation Practice, 17,* 121–130.

Chen, H. T., & Rossi, P. H. (1980). The multi-goals, theory-driven approach to evaluation: A model linking basic and applied social sciences. *Social Forces, 59,* 106–122.

Chen, H.T., & Rossi, P. H. (1983). Evaluating with sense: The theory-driven approach. *Evaluation Review, 7,* 283–302.

Chinman, M., Hunter, S. B., Ebener, P., Paddock, S., Stillman, L., Imm, P., & Wanderman, A. (2008). The Getting To Outcomes demonstration and evaluation: An illustration of the prevention support system. *American Journal of Community Psychology, 41*(3–4), 206–224.

Christie, C. A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. In C. A. Christie (Ed.), *The practice-theory relationship in evaluation.* New Directions for Evaluation, No. 9, 7–36. San Francisco: Jossey-Bass.

Christie, C. A., & Barela, E. (2008). Internal evaluation in a large urban school district: A Title I best practices study. *American Journal of Evaluation, 29*(4), 531–546.

Christie, C., & Conner, R. F. (2005). A conversation with Ross Conner: The Colorado Trust Community-Based Collaborative Evaluation. *American Journal of Evaluation*, *26*, 369–377.

Christie, C., & Preskill, H. (2006). Appreciative Inquiry as a method for evaluation: An interview with Hallie Preskill. *American Journal of Evaluation*, *27*(4), 466–474.

Chubb, J. E., & Moe, T. M. (1990). *Politics, markets, and America's schools.* Washington, DC: Brookings Institute.

Compton, D. W., & Baizerman, M. (2007). Defining evaluation capacity building. *American Journal of Evaluation*, *28*(2), 118–119.

Compton, D.W., Baizerman, M., & Stockdill, S. H. (Eds.). (2002). *The art, craft, and science of evaluation capacity building.* New Directions for Evaluation. San Francisco: Jossey-Bass.

Compton, D. W., Glover-Kudon, R., Smith, I. E., & Avery, M. E. (2002). Ongoing capacity building in the American Cancer Society (ACS) 1995–2001. In D. W. Compton, M. Braizerman, & S. H. Stockdill (Eds.), *The art, craft, and science of evaluation capacity building.* New Directions for Evaluation, No. 93, 47–62. San Francisco: Jossey-Bass.

Converse, P. D., Wolfe, E. W., Huang, X., & Oswald, F. L. (2008). Response rates for mixed-mode surveys using mail and e-mail/web. *American Journal of Evaluation*, *29*(1), 99–107.

Cook, D. L. (1966). *Program evaluation and review technique: Applications in education* (Monograph no. 17). Washington, DC: U. S. Office of Education Cooperative Research.

Cook, T. D. (1997). Lessons learned in evaluation over the past 25 years. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook.* Thousand Oaks, CA: Sage.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Skokie, IL; Rand McNally.

Cooksy, L. J. (2000). Commentary: Auditing the off-the-record case. *American Journal of Evaluation*, *21*, 122–128.

Cousins, J. B. (2003). Utilization effects of participatory evaluation. In T. Kellaghan, D. L. Stufflebeam, & L. A. Wingate (Eds.), *International handbook of educational evaluation* (pp. 245–266). Dordrecht, The Netherlands: Kluwer Academic.

Cousins, J. B. (2005). Will the real empowerment evaluation please stand up? A critical friend perspective. In D. M. Fetterman & A. Wandersman (Eds.), *Empowerment evaluation principles in practice*, pp. 183–208. New York: Guilford.

Cousins, J. B., Donohue, J. J., & Bloom, G. A. (1996). Collaborative evaluation in North America: Evaluators' self-reported opinions, practices, and consequences. *Evaluation Practice*, *17*(3), 207–226.

Cousins, J. B., & Earl, L. M. (1992). The case for participatory evaluation. *Educational Evaluation and Policy Analysis*, *14*(4), 397–418.

Cousins, J. B., & Earl, L. M. (Eds.). (1995). *Participatory evaluation in education: Studies in evaluation use and organizational learning*. London: Falmer.

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research*, *56*, 331–364.

Cousins, J. B., & Shula, L. M. (2006). A comparative analysis of evaluation utilization and its cognate fields of inquiry: Current issues and trends. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation* (pp. 266–291). Thousand Oaks, CA: Sage.

Cousins, J. B., & Shula, L. M. (2008). Complexities in setting standards in collaborative evaluation. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 139–158). New York: Guilford.

Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. In E. Whitmore (Ed.), *Understanding and practicing participatory evaluation.* New Directions for Evaluation, No. 80, 5–23. San Francisco: Jossey-Bass.

Covert, R. W. (1992, November). *Successful competencies in preparing professional evaluators*. Paper presented at the annual meeting of the American Evaluation Association, Seattle, WA.

Cox, J. (2007). *Your opinion please: How to build the best questionnaires in the field of education* (2nd ed.). Thousand Oaks, CA: Sage.

Crary, D. (2010, February 4). Child-abuse crackdown makes huge dent in cases. *Denver Post*, pp. 1, 9a.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record*, *64*, 672–683.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. (1980). *Toward reform of program evaluation.* San Francisco: Jossey-Bass.

Cullen, A. (2009). *The politics and consequences of stakeholder participation on international development evaluation.* Unpublished doctoral dissertation, Western Michigan University.

Dahler-Larsen, P. (2003). The political in evaluation. *E-journal Studies in Educational Policy and Educational Philosophy, 1*. http://www.upi.artisan.se.

Dahler-Larsen, P. (2006). Evaluation after disenchantment? Five issues shaping the role of evaluation in society. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation.* Thousand Oaks, CA: Sage.

Datta, L. (1999). The ethics of evaluation neutrality and advocacy. In J. L. Fitzpatrick & M. Morris (Eds.), *Current and emerging ethical challenges in evaluations.* New Directions for Evaluation, No. 82, 77–88. San Francisco: Jossey-Bass.

Datta, L. (2006). The practice of evaluation: Challenges and new directions. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation.* Thousand Oaks, CA: Sage.

Datta, L. (2007). Looking at the evidence: What variations in practice might indicate. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation methodology: Building the evidence base for method choice in government sponsored evaluation.* New Directions for Evaluation, No. 113, 35–54. San Francisco: Jossey-Bass.

Datta, L. E., & Miller, R. (2004). The oral history of evaluation Part II: The professional development of Lois-Ellin Datta. *American Journal of Evaluation, 25*, 243–253.

Davidson, J. (2005). Criteria. In S. Mathison (Ed.), *Encyclopedia of evaluation*, p. 91. Thousand Oaks, CA: Sage.

Davidson, J. (2010). Posting on evaluation discussion. American Evaluation Association. Thought leaders. 1/20/2010.

deLeeuw, E. D., Hox, J. J., & Dillman, D. A. (2008). *International handbook of survey methodology.* Psychology Press.

deLeon, L., & Denhardt, R. (2000). The political theory of reinvention. *Public Administration Review, 60*(1), 89–97.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2005). *The Sage handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Dewey, J. D., Montrosse, B. E., Shroter, D. C., Sullins, C. D., & Mattox, J. R. (2008). Evaluator competencies: What's taught versus what's sought. *American Journal of Evaluation, 29,* 268–287.

Dickeson, R. C. (2006). The need for accreditation reform. Retrieved from http://www.ed.gov/about/bdscomm/list/hiedfuture/reports/dickeson.pdf. June 1, 2010.

Dillman, D. A., Smyth, J., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: John Wiley.

Dillman, D. A., & Tarnai, J. (1991). Mode effects of cognitively-designed recall questions: A comparison of answers to telephone and mail surveys. In P. P. Biemer, R. M. Brover, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys*. New York: Wiley.

Dodson, S. C. (1994). *Interim summative evaluation: Assessing the value of a long term or ongoing program, during its operation.* Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Donaldson, S. E., & Scriven, M. (Eds.). (2003). *Evaluating social programs and problems: Visions for the new millenniums.* Mahwah, NJ: Erlbaum.

Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications.* New York: Erlbaum.

Donmoyer, R. (2005). Connoisseurship. In S. Matthison (Ed.), *Encyclopedia of evaluation.* Thousand Oaks, CA: Sage.

Duignan, P. (2003). Mainstreaming evaluation or building evaluation capability? Three key elements. In J. J. Barnette & J. R. Sanders (Eds.), *The mainstreaming of evaluation.* New Directions for Evaluation, No. 99, pp. 7–22. San Francisco: Jossey-Bass.

Eden, C., & Ackermann, F. (1998). *Making strategy: The journey of strategic management.* London: Sage.

Eisner, E. W. (1975, April). *The perceptive eye: Toward the reformation of educational evaluation.* Invited address at the American Educational Research Association, Washington, DC.

Eisner, E. W. (1976). Educational connoisseurship and criticism: Their form and function in educational evaluation. *Journal of Aesthetic Education, 10*, 135–150.

Eisner, E. W. (1979a). *The educational imagination: On the design and evaluation of school programs.* New York: Macmillan.

Eisner, E. W. (1979b). The use of qualitative forms of evaluation for improving educational practice. *Educational Evaluation and Policy Analysis, 1*, 11–19.

Eisner, E. W. (1985). *The art of educational evaluation: A personal view*. Philadelphia, PA: The Falmer Press.

Eisner, E. W. (1991a). Taking a second look: Educational connoisseurship revisited. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century.* Ninetieth Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press.

Eisner, E. W. (1991b). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice.* New York: Macmillan.

Eisner, E. W. (2004). The roots of connoisseurship and criticism: A personal journey. In M. C. Alkin (Ed.), *Evaluation roots.* Thousand Oaks, CA: Sage.

Elliott, J. (2005). Action research. In S. Mathison (Ed.), *Encyclopedia of evaluation.* Thousand Oaks, CA: Sage.

Evaluation Research Society Standards Committee. (1982). Evaluation Research Society standards for program evaluation. In P. H. Rossi (Ed.), *Standards for evaluation practice.* New Directions for Program Evaluation, No. 15, 7–19. San Francisco: Jossey-Bass.

Fals-Borda, O., & Anisur-Rahman, M. (1991). *Action and knowledge: Breaking the monopoly with participatory action research*. New York: Apex Press.

Fetterman, D. M. (1984). *Ethnography in educational evaluation.* Beverly Hills, CA: Sage.

Fetterman, D. M. (1994). Empowerment evaluation. 1993 presidential address. *Evaluation Practice, 15*, 1–15.

Fetterman, D. M. (1996). Empowerment evaluation: An introduction to theory and practice. In D. M. Fetterman, S. Kaftarian, & A. Wandersman (Eds.), *Empowerment evaluation: Knowledge and tools for self-assessment and accountability.* Thousand Oaks, CA: Sage.

Fetterman, D. M. (2001a). *Foundations of empowerment evaluation.* Thousand Oaks, CA: Sage.

Fetterman, D. M. (2001b). The transformation of evaluation into a collaboration: A vision of evaluation in the 21st century. *American Journal of Evaluation, 22*(3), 381–385.

Fetterman, D. M. (2005). In response to Drs. Patton and Scriven. *American Journal of Evaluation, 26*(3), 418–420.

Fetterman, D. M., Kaftarian, S., and Wandersman, A. (Eds.). (1996). *Empowerment evaluation: Knowledge and tools for self-assessment and accountability*. Thousand Oaks, CA: Sage.

Fetterman, D. M., & Wandersman, A. (Eds.). (2005). *Empowerment evaluation principles in practice.* New York: Guilford.

Fetterman, D. M., & Wandersman, A. (2007). Empowerment evaluation: Yesterday, today, and tomorrow. *American Journal of Evaluation, 28*(2), 179–198.

Fink, A. (2008). *How to conduct surveys: A step-by-step guide* (4th ed.). Thousand Oaks, CA: Sage.

Fischer, C. T., & Wertz, F. J. (2002). Empirical phenomenological analyses of being criminally victimized. In A. M. Huberman & M. B. Miles (Eds.), *The qualitative researcher's companion.* Thousand Oaks, CA: Sage.

Fitzpatrick, J. L. (1988). *Alcohol education programs for drunk drivers.* Colorado Springs: Center for Community Development and Design.

Fitzpatrick, J. L. (1989). The politics of evaluation with privatized programs: Who is the audience? *Evaluation Review, 13*, 563–578.

Fitzpatrick, J. L. (1992). Problems in the evaluation of treatment programs for drunk drivers: Goals and outcomes. *Journal of Drug Issues, 22*, 155–167.

Fitzpatrick, J. L. (1994). Alternative models for the structuring of professional preparation programs. In J. W. Altschuld & M. Engle (Eds.), *The preparation of professional evaluators: Issues, perspectives, and programs.* New Directions for Program Evaluation, No. 62, 41–50. San Francisco: Jossey-Bass.

Fitzpatrick, J. L. (1999). Ethics in disciplines and professions related to evaluation. In J. L. Fitzpatrick & M. Morris (Eds.), *Current and emerging ethical challenges in evaluation.* New Directions for Evaluation, No. 82. San Francisco: Jossey-Bass.

Fitzpatrick, J. L. (2004). Exemplars as case studies: Reflections on the link between theory, practice, and context. *American Journal of Evaluation, 25*, 541–559.

Fitzpatrick, J. L. (2005). Informed consent. In S. Mathison (Ed.), *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.

Fitzpatrick, J. L., & Bickman, L. (2002). Evaluation of the Ft. Bragg and Stark County systems of care for children and adolescents: A dialogue with Len Bickman. *American Journal of Evaluation, 23*, 67–80.

Fitzpatrick, J. L., & Bledsoe, K. (2007). Evaluation of the Fun with Books Program: A dialogue with Katrina Bledsoe. *American Journal of Evaluation, 28*, 522–535.

Fitzpatrick, J. L., & Donaldson, S. I. (2002). Evaluation of the Work and Health Initiative: A dialogue with Stewart Donaldson. *American Journal of Evaluation, 23*, 347–365.

Fitzpatrick, J. L., & Fetterman, D. (2000). The evaluation of the Stanford Teacher Education Program (STEP): A dialogue with David Fetterman. *American Journal of Evaluation, 20*, 240–259.

Fitzpatrick, J. L., & Greene, J. (2001). Evaluation of the Natural Resources Leadership Program: A dialogue with Jennifer Greene. *American Journal of Evaluation, 22*, 81–96.

Fitzpatrick, J. L., & Henry, G. (2000). The Georgia Council for School Performance and its performance monitoring system: A dialogue with

Gary Henry. *American Journal of Evaluation*, *21*, 105–117.

Fitzpatrick, J. L., & King, J. A. (2009). Evaluation of the special education program at the Anoka-Hennepin School District: An interview with Jean A. King. In J. Fitzpatrick, C. Christie, & M. M. Mark (Eds.), *Evaluation in action*. Thousand Oaks, CA: Sage.

Fitzpatrick, J. L., & Miller-Stevens, K. (2009). A case study of measuring outcomes in an MPA program. *Journal of Public Affairs Education*, *15*(1), 17–31.

Fitzpatrick, J. L., & Riccio, J. A. (1997). A dialogue about an award-winning evaluation of GAIN: A welfare-to-work program. *Evaluation Practice*, *18*, 241–252.

Fitzpatrick, J. L., & Rog, D. J. (1999). The evaluation of the Homeless Families Program. A dialogue with Debra Rog. *American Journal of Evaluation*, *20*, 562–575.

Flexner, A. (1910). *Medical education in the United States and Canada* (Bulletin no. 4). New York: Carnegie Foundation for the Advancement of Teaching.

Flexner, A. (1960). *Abraham Flexner: An autobiography*. New York: Simon & Schuster.

Floden, R. E. (1983). Flexner, accreditation, and evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation*. Boston: Kluwer-Nijhoff.

Fontana, A., & Frey, J. H. (2000). Interviewing: From structured questions to negotiated text. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research*. Thousand Oaks, CA: Sage.

Frankfort-Nachmias, C., & Nachmias, D. (2008). *Research methods in the social sciences* (7th ed.). New York: Worth Publishers.

Frechtling, J. (2002). *The 2002 user-friendly handbook for project evaluation*. Arlington, VA: National Science Foundation.

Freire, P. (1970). *Pedagogy of the oppressed.* New York: Seabury Press.

Freire, P. (1982). Creating alternative research methods: Learning to do it by doing it. In B. Hall, A. Gillette, & R. Tandon (Eds.), *Creating knowledge: A monopoly. Participatory research in development.* New Delhi: Society for Participatory Research in Asia.

Fricker, R. D., & Schonlau, M. (2002). Advantages and disadvantages of Internet research surveys: Evidence from the literature. *Field Methods*, *14*, 347–367.

Frierson, H., Hood, S., & Hughes, S. (2002). Strategies that address culturally responsive evaluation. In J. Frechtling (Ed.), *The 2002 user-friendly handbook for project evaluation*. Arlington, VA: National Science Foundation.

Gephart, W. J. (1978). *The facets of the evaluation process: A starter set.* Unpublished manuscript. Bloomington, IN: Phi Delta Kappan.

Gilmour, J. B., & Davis, D. E. (2006). Does performance budgeting work? An examination of the Office of Management and Budget's PART scores. *Public Administration Review*, *66*, 742–752.

Girard, G., & Impara, J. C. (2005). Making the cut: The cut score setting process in a public school district. *Applied Measurement in Education*, *18*(3), 289–232.

Gitomar, D. (2007). *Teacher quality in a changing policy landscape: Improvements in the teacher pool*. Princeton, NJ: Educational Testing Service.

Godfrey-Smith, P. (2003). *An introduction to the philosophy of science*. Chicago: University of Chicago Press.

Goldring, E. B., & Shapira, R. (1993). Choice, empowerment, and involvement: What satisfies parents? *Educational Evaluation and Policy Analysis*, *15*, 396–409.

Goodlad, J. (1979). *What schools are for*. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Gore, A. (1993). *From red tape to results: Creating a government that works better and costs less: The report of the National Performance Review*. New York: Plume.

Greene, J. C. (1987). Stakeholder participation in evaluation design: Is it worth the effort? *Evaluation and Program Planning*, *10*, 379–394.

Greene, J. C. (1988). Stakeholder participation and utilization in program evaluation. *Evaluation Review*, *12*, 91–116.

Greene, J. C. (1997). Evaluation as advocacy. *Evaluation Practice*, *18*, 25–35.

Greene, J. C. (2000). Challenges in practicing deliberative democratic evaluation. In K. E. Ryan and L. DeStefano (Eds.), *Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation.* New Directions for Evaluation, No. 85, pp. 13–26. San Francisco: Jossey-Bass.

Greene, J. C. (2005). Mixed methods. In S. Mathison (Ed.), *Encyclopedia of evaluation.* Thousand Oaks, CA: Sage.

Greene, J. C. (2006). Evaluation, democracy, and social change. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation*. London: Sage Publications.

Greene, J. C. (2008). Memories of a novice, learning from a master. *American Journal of Evaluation*, *29*(3), 322–324.

Greene, J. C., & Abma, T. (Eds.). (2001). *Responsive evaluation.* New Directions for Evaluation, No. 92. San Francisco: Jossey-Bass.

Greene, J. C., & Caracelli, V. J. (Eds.). (1997). *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms.* New Directions for Program Evaluation, No. 74. San Francisco: Jossey-Bass.

Greene, J. C., & Henry, G. T. (2005). Qualitative-quantitative debate in evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.

Greene, J. C., Lipsey, M. W., Schwandt, T. A., Smith, N. L., & Tharp, R. G. (2007). Method choice: Five discussant commentaries. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation methodology: Building the evidence based for method choice in government sponsored evaluation.* New Directions for Evaluation, No. 113, 111–128. San Francisco: Jossey-Bass.

Greiner, J. M. (2004). Trained observer ratings. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (2nd ed.). San Francisco: Jossey-Bass.

Guba, E. G. (1969). The failure of educational evaluation. *Educational Technology*, *9*, 29–38.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation*. San Francisco: Jossey-Bass.

Guba, E. G., & Lincoln, Y. S. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Thousand Oaks, CA: Sage.

Guba, E., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research.* Thousand Oaks, CA: Sage.

Hall, B. L. (1992). From margins to center? The development and purpose of participatory research. *American Sociologist, 3,* 15–28.

Hammond, R. L. (1973). Evaluation at the local level. In B. R. Worthen & J. R. Sanders, *Educational evaluation: Theory and practice*. Belmont, CA: Wadsworth.

Hanssen, C. E., Lawrenz, F., & Dunet, D. O. (2008). Concurrent meta-evaluation: A critique. *American Journal of Evaluation*, *29*(4), 572–582.

Hebert, Y. M. (1986). Naturalistic evaluation in practice: A case study. In D. D. Williams (Ed.), *Naturalistic evaluation.* New Directions for Program Evaluation, No. 30, 3–21. San Francisco: Jossey-Bass.

Hendricks, M., & Conner, R. F. (1995). International perspectives on the Guiding Principles. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.), *Guiding principles for evaluators*. New Directions for Program Evaluation, No. 66, 77–90. San Francisco: Jossey-Bass.

Hendricks, M., Plantz, M. C., & Pritchard, K. J. (2008). Measuring outcomes of United Way-funded programs: Expectations and reality. In J. G. Carman & K. A. Fredericks (Eds.), *Nonprofits and evaluation*. New Directions for Evaluation, No. 119, 13–35. San Francisco: Jossey-Bass.

Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.

Henry, G. T. (1996). Does the public have a role in evaluation? Surveys and democratic discourse. In M. T. Braverman & J. K. Slater (Eds.), *Advances in survey research*. New Directions for Program Evaluation, No. 70, 3–16. San Francisco: Jossey-Bass.

Henry, G. T. (Ed.). (1997). *Creating effective graphs: Solutions for a variety of evaluation data.* New Directions for Program Evaluation, No. 73. San Francisco: Jossey-Bass.

Henry, G. T. (2000). Why not use? In V. J. Caracelli & H. Preskill (Eds.), *The expanding scope of evaluation use.* New Directions for Evaluation, No. 88, 85–98. San Francisco: Jossey-Bass.

Henry, G. T., & Mark, M. M. (2003). Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation*, *24*(3), 293–314.

Heron, J. (1981). Validity in co-operative inquiry. In P. Reason (Ed.), *Human inquiry in action.* London: Sage.

Hodgkinson, H., Hurst, J., & Levine, H. (1975). *Improving and assessing performance: Evaluation in higher education*. Berkeley, CA: University of California Center for Research and Development in Higher Education.

Honea, G. E. (1992*). Ethics and public sector evaluators: Nine case studies.* Unpublished doctoral dissertation. University of Virginia, Department of Educational Studies.

Hood, S. L. (2000). Commentary on deliberative democratic evaluation. In K. E. Ryan & L. Destefano (Eds.), *Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation*. New Directions for Evaluation, No. 83. San Francisco: Jossey-Bass.

Hood, S. L. (2001). Nobody knows my name: In praise of African American evaluators who were responsive. In J. C. Greene & T. A. Abma (Eds.), *Responsive evaluation*. New Direction for Evaluation, No. 92. San Francisco: Jossey-Bass.

Hood, S. L. (2005). Culturally responsive evaluation. In S. Mathison (Ed.), *The encyclopedia of evaluation.* Thousand Oaks, CA: Sage.

Horst, P., Nay, J. N., Scanlon, J. W., & Wholey, J. S. (1974). Program management and the federal evaluator. *Public Administration Review, 34,* 300–308.

House, E. R. (1980). *Evaluating with validity.* Beverly Hills, CA: Sage.

House, E. R. (1990). Methodology and justice. In K. A. Sirotnik (Ed.), *Evaluation and social justice: Issues in public education*. New Directions for Program Evaluation, No. 45, 23–36. San Francisco: Jossey-Bass.

House, E. R. (1993). *Professional evaluation.* Newbury Park, CA: Sage.

House, E. R. (2001a). Responsive evaluation (and its influence on deliberative democratic evaluation). In J. C. Greene & T. A. Abma (Eds.), *Responsive evaluation.* New Directions for Evaluation, No. 92, 23–30. San Francisco: Jossey-Bass.

House, E. R. (2001b). Unfinished business: Causes and values. *American Journal of Evaluation, 22,* 309–316.

House, E. R., & Howe, K. R. (1999). *Values in evaluation and social research*. Thousand Oaks, CA: Sage.

House, E. R., & Howe, K. R. (2000). Deliberative democratic evaluation. In K. E. Ryan & L. DeStefano (Eds.*), Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation*. New Directions for Evaluation, No. 85, 3–12. San Francisco: Jossey-Bass.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis (or dogmas die hard). *Educational Researcher, 17*(8), 10–16.

Hoxby, C. M. (2000). Does competition among public schools benefit students and taxpayers? *American Economic Review, 90,* 1209–1238.

Huberman, A. M., & Cox, P. L. (1990). Evaluation utilization: Building links between action and reflections. *Studies in Educational Evaluation, 16,* 157–179.

Huberman, A. M., & Miles, M. B. (2002). *The qualitative researcher's companion.* Thousand Oaks, CA: Sage.

Humphreys, L. (1975). *Tearoom trade: Impersonal sex in public places*. Chicago: Aldine.

Hurworth, R. (2005). Document analysis. In S. Mathison (Ed.), *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.

Huxley, E. (1982). *The flame trees of Thika: Memories of an African childhood.* London: Chatto & Windus.

International Organization for Coordination in Evaluation. (2003). *Mission statement*. Downloaded on March 8, 2009 from http://www.internationalevaluation.com/overview/mission-vision.shtml.

Jaeger, R. M. (1990). *Statistics: A spectator sport* (2nd ed.). Newbury Park, CA: Sage.

Johnson, E. C., Kirkhart, K. E., Madison, A. M., Noley, G. B., & Solano-Flores, G. (2008). The impact of narrow views of scientific rigor on evaluation practices for underrepresented groups. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation*. New York: Guilford Press.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials*. New York: McGraw-Hill.

Joint Committee on Standards for Educational Evaluation. (1988). *Personnel evaluation standards.* Newbury Park, CA: Corwin.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (2010). *The program evaluation standards* (3rd ed.). Thousand Oaks, CA: Sage.

Jorgensen, D. L. (1989). *Participant observation: A methodology for human studies*. Newbury Park, CA: Sage.

Kahn, J. P., & Mastroianni, A. P. (2001). Doing research well by doing it right. *Chronicle of Higher Education*, 15 February, B24.

Kane, M. (1995). Examinee-centered vs. task-centered standard setting. In *Proceedings of the Joint Conference on Standard Setting in Large-Scale Assessment* (Vol. 2, pp. 119–139). Washington, DC: National Assessment Governing Boards and National Center for Education Statistics.

Kaplan, A. (1964). *The conduct of inquiry*. San Francisco: Chandler.

Karlsson, O. (1998). Socratic dialogue in the Swedish political context. In T. A. Schwandt (Ed.), *Scandinavian perspectives on the evaluator's role in informing social policy*. New Directions for Evaluation, No. 77. San Francisco: Jossey-Bass.

Kee, J. E. (2004). Cost-effectiveness and cost-benefit analysis. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation*. San Francisco: Jossey-Bass.

Kellow, J. T. (1998). Beyond statistical significance tests: The importance of using other estimates of treatment effects to interpret evaluation results. *American Journal of Evaluation, 19,* 123–134.

King, J. A. (1998). Making sense of participatory evaluation practice. In E. Whitmore (Ed.), *Understanding and practicing participatory*

evaluation. New Directions for Evaluation, No. 80, 57–68. San Francisco: Jossey-Bass.

King, J. A. (2002). Building the evaluation capacity of a school district. In D. W. Compton, M. Braizerman, & S. H. Stockdill (Eds.), *The art, craft, and science of evaluation capacity building.* New Directions for Evaluation, No. 93, 63–80. San Francisco: Jossey-Bass.

King, J. A. (2005). Participatory evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.

King, J. A. (2007). Developing evaluation capacity through process use. In J. B. Cousins (Ed.), *Process use in theory, research, and practice.* New Directions for Evaluation, No, 116, 45–59. San Francisco: Jossey-Bass.

King, J. A., Stevahn, L., Ghere, G., & Minnema, J. (2001). Toward a taxonomy of essential evaluator competencies. *American Journal of Evaluation*, *22*, 229–247.

Kirkhart, K. E. (1995). Seeking multicultural validity: A postcard from the road. *American Journal of Evaluation*, *16*, 1–12.

Kirkhart, K. E. (2000). Reconceptualizing evaluation use: An integrated theory of influence. In V. J. Caracelli & H. Preskill (Eds.), *The expanding scope of evaluation use*. New Directions for Evaluation, No. 88. San Francisco: Jossey-Bass.

Kirkpatrick, D. L. (1977). Evaluating training programs: Evidence vs. proof. *Training and Development Journal*, *31*, 9–12.

Kirkpatrick, D. L. (1983). Four steps to measuring training effectiveness. *Personnel Administrator*, *28*, 19–25.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels* (3rd ed.). San Francisco: Berrett-Koehler Publishers.

Kline, R. B. (2009). *Becoming a behavioral science researcher: A guide to producing research that matters*. New York: Guilford Press.

Knowlton, L. W., & Phillips, C. C. (2009). *The logic model guidebook.* Thousand Oaks, CA: Sage.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: Handbook II.* Affective domain. New York: David McKay.

Krueger, R. A. (2005). Focus groups. In S. Mathison (Ed.), *Encyclopedia of evaluation.* Thousand Oaks, CA: Sage.

Krueger, R. A., & Casey, M. (2009). *Focus groups: A practical guide for applied research* (4th ed.). Thousand Oaks, CA: Sage.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Kushner, S. (2000). *Personalizing evaluation.* London: Sage.

Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing.* Thousand Oaks, CA: Sage.

Lambur, M. T. (2008). Organizational structures that support internal program evaluation. In M. T. Braverman, M. Engle, M. E. Arnold, & R. A. Rennekamp (Eds.), *Program evaluation in a complex organizational system: Lessons from Cooperative Extension*. New Directions for Evaluation, No. 120, 41–54. San Francisco: Jossey-Bass.

LaVelle, J. M., & Donaldson, S. I. (2010). University-based evaluation training programs in the United States: 1980–2008: An empirical examination. *American Journal of Evaluation*, *31*(1), 9–23.

Layard, R., & Glaister, S. (Eds.). (1994). *Cost-benefit analysis*. New York: Cambridge University Press.

Lee, A. M., & Holly, F. R. (1978, April). *Communicating evaluation information: Some practical tips that work.* Paper presented at the meeting of the American Educational Research Association, Toronto.

Leeuw, F. L. (2003). Reconstructing program theories: Methods available and problems to be solved. *American Journal of Evaluation, 24*(1), 5–20.

Levin, H. M. (1986). A benefit-cost analysis of nutritional programs for anemia reduction. *World Bank Research Observer, 1*(2), 219–245.

Levin, H. M. (2005). Cost-benefit analysis. In S. Mathison (Ed.), *Encyclopedia of evaluation.* Thousand Oaks, CA: Sage.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Leviton, L. C. (2001). Presidential address: Building evaluation's collective capacity. *American Journal of Evaluation*, *22*(1), 1–12.

Leviton, L. C. (2003). Evaluation use: Advances, challenges, and applications. *American Journal of Evaluation, 24*(4), 525–535.

Leviton, L. C., & Hughes, E. F. X. (1981). Research on the utilization of evaluations: A review and synthesis. *Evaluation Review*, *5*, 524–548.

Lewin, K. (1946). Action research and minority problems. In K. Lewin (Ed.), *Resolving social conflicts: Selected papers on group dynamics*. New York: Harper & Row.

Light, R. J. (Ed.). (1983). *Evaluation studies review annual* (Vol. 8). Beverly Hills, CA: Sage.

Light, R. J., & Smith, P. V. (1970). Choosing a future: Strategies for designing and evaluating new programs. *Harvard Educational Review*, *40*, 1–28.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Beverly Hills, CA: Sage.

Lincoln, Y. S., & Guba, E. G. (2004). The roots of fourth generation evaluation: Theoretical and methodological origins. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences.* Thousand Oaks, CA: Sage.

Lindblom, C. E. (1977). *Politics and markets: The world's political economic systems*. New York: Basic Books.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Lipsey, M. W. (1993). Theory as method: Small theories of treatments. In L. B. Sechrest & A. G. Scott (Eds.), *Understanding causes and generalizing about them*. New Directions for Evaluation, No. 57, 5–38. San Francisco: Jossey-Bass.

Lipsey, M. W. (2000). Method and rationality are not social diseases. *American Journal of Evaluation, 21*, 221–224.

Lipsey, M. W. (2007). Method choice for government evaluation: The beam in our own eye. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation methodology: Building the evidence base for method choice in government sponsored evaluation*. New Directions for Evaluation, No. 113, 113–115. San Francisco: Jossey-Bass.

Love, A. J. (1983). The organizational context and the development of internal evaluation. In A. J. Love (Ed.), *Developing effective internal evaluation*. New Directions for Program Evaluation, No. 20, 5–22. San Francisco: Jossey-Bass.

Love, A. J. (1991). *Internal evaluation: Building organizations from within.* Newbury Park, CA: Sage.

Lovell, R. G. (1995). Ethics and internal evaluators. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye, C. (Eds.), *Guiding principles for evaluators.* New Directions for Program Evaluation, No. 66, 61–67. San Francisco: Jossey-Bass.

Lyman, D. R., Milich, R., Zimmerman, R., Novak, S. P., Logan, T. K., & Martin, C. (1999). Project DARE: No effects at 10-year follow-up. *Journal of Consulting and Clinical Psychology, 67*, 590–593.

Mabry, L. (1999). Circumstantial ethics. *American Journal of Evaluation, 20*, 199–212.

Mabry, L. (2005). Portfolio. In S. Mathison (Ed.), *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.

MacDonald, B. (1974). Evaluation and the control of education. In B. MacDonald & R. Walker (Eds.), *Innovation, evaluation, research, and the problem of control*. (SAFARI ONE). Norwich, UK: Centre for Applied Research in Education, University of East Anglia.

MacDonald, B. (1976). Evaluation and the control of education. In D. Tawney (Ed.), *Curriculum evaluation today: Trends and implications.* Schools Council Research Studies Series. London: Macmillan.

Mackay, K. (2002). The World Bank's ECB experience. In D. W. Compton, M. Braizerman, & S. H. Stockdill (Eds.), *The art, craft, and science of evaluation capacity building.* New Directions for Evaluation, No. 93, 81–100. San Francisco: Jossey-Bass.

MacNeil, C. (2000). Surfacing the realpolitik: Democratic evaluation in an antidemocratic environment. In K. E. Ryan and L. DeStefano (Eds.), *Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation*. New Directions for Evaluation, No. 85, 51–62. San Francisco: Jossey-Bass.

Madaus, G. F. (2004). Ralph W. Tyler's contribution to program evaluation. In M. C. Alkin (Ed.), *Evaluation roogs: Tracing theorists' views and influences.* Thousand Oaks, CA: Sage.

Madaus, G. F., & Stufflebeam, D. L. (Eds.). (1989). *Educational evaluation: Classic works of Ralph W. Tyler*. Boston: Kluwer Academic.

Madison, A. (2007). New Directions for Evaluation coverage of cultural issues and issues of significance to underrepresented groups. In S. Mathison (Ed.), *Enduring issues in evaluation: The 20th anniversary of the collaboration between NDE and AEA*. New Directions for Evaluation, No. 114, 107–114. San Francisco: Jossey-Bass.

Mark, M. M. (2003). Program evaluation. In S. A. Schinka & W. Velicer (Eds.), *Comprehensive handbook of psychology* (Vol 2, pp. 323–347). New York: Wiley.

Mark, M. M. (2007). AEA and evaluation: 2006 (and beyond). In S. Mathison (Ed.), *Enduring issues in evaluation: The 20th anniversary of the collaboration between NDE and AEA.* New Directions for Evaluation, No. 114, 115–119. San Francisco: Jossey-Bass.

Mark, M. M., & Henry, G. T. (2006). Methods for policy-making and knowledge development evaluations. In E. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation* (pp. 317–339). Thousand Oaks, CA: Sage.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). E*valuation: An integrated framework for understanding, guiding, and improving policies and programs.* San Francisco: Jossey-Bass.

Mark, M. M., & Shotland, R. L. (1985). Stakeholder-based evaluation and value judgments: The role of perceived power and legitimacy in the selection of stakeholder groups. *Evaluation Review, 9*, 605–626.

Mathison, S. (1999). Rights, responsibilities, and duties: A comparison of ethics for internal and external evaluators. In J. L. Fitzpatrick & M. Morris (Eds.), *Current and emerging ethical challenges in evaluations.* New Directions for Evaluation, No. 82, 25–34. San Francisco: Jossey-Bass.

Mathison, S. (2007). 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*. New York: Guilford Press.

McCall, R. B., Ryan, C. S., & Green, B. L. (1999). Some non-randomized constructed comparison groups for evaluating age-related outcomes of intervention programs. *American Journal of Evaluation*, *2*, 213–226.

McClintock, C. (1987). Conceptual and action heuristics: Tools for the evaluator. In L. Bickman (Ed.), *Using program theory in evaluation*. New Directions for Program Evaluation, No. 33, 43–57. San Francisco: Jossey-Bass.

McCracken, G. (1988). *The long interview*. Qualitative Research Methods Series, no. 13. Newbury Park, CA: Sage.

McLaughlin, J. A., & Jordan, G. B. (2004). Using logic models. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (2nd ed.). San Francisco: Jossey-Bass.

McTaggert, R. (1991). Principles for participatory action research. *Adult Education Quarterly, 41*(3), 168–187.

Merriam-Webster's Dictionary (2009). http://www.merriam-wester.com/dictionary/politics.

Mertens, D. M. (1994). Training evaluators: Unique skills and knowledge. In J. W. Altschuld & M. Engle (Eds.), *The preparation of professional evaluators: Issues, perspectives, and programs*. New Directions for Program Evaluation, No. 62, 17–27. San Francisco: Jossey-Bass.

Mertens, D. M. (1999). Inclusive evaluation: Implications of transformative theory for evaluation. *American Journal of Evaluation*, *20*, 1–14.

Mertens, D. M. (2001). Inclusivity and transformation: Evaluation in 2010. *American Journal of Evaluation*, *22*, 367–374.

Mertens, D. M. (2008). *Transformative research and evaluation*. New York: Guilford Press.

Milanowski, A. (2008). *How to pay teachers for student performance outcomes.* Consortium for Policy Research in Education (CPRE) Project/Strategic Management of Human Capital. Retrieved March 8, 2010 from http://www.smhc-cpre.org/wp-content/uploads/2008/10/cb-paper-4-paying-for-student-performance.pdf.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Milgram, S. (1974). *Obedience and authority.* New York: Harper & Row.

Miller, R. L., & Campbell, R. (2006). Taking stock of empowerment evaluation: An empirical review. *American Journal of Evaluation*, *27*(3), 296–319.

Mills, A. S., Massey, J. G., & Gregersen, H. M. (1980). Benefit-cost analysis of Voyageurs National Park. *Evaluation Review*, *4*, 715–738.

Milstein, B., Chapel, T. J., Wetterhall, S. F., & Cotton, D. A. (2002). Building capacity for program evaluation at the Centers for Disease Control and Prevention. In D. W. Compton, M. Braizerman, & S. H. Stockdill (Eds.), *The art, craft, and science of evaluation capacity building.* New Directions for Evaluation, No. 93, 27–46. San Francisco: Jossey-Bass.

Morell, J. A. (2005). Product evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation.* Thousand Oaks, CA: Sage.

Morris, M. (Ed.). (2008). *Evaluation ethics for best practice: Cases and commentary.* New York: Guilford Press.

Morris, M., & Cohn, R. (1993). Program evaluators and ethical challenges: A national survey. *Evaluation Review*, *17*, 621–642.

Morris, M., Cooksy, L. J., & Knott, T. D. (2000). The off-the-record case. *American Journal of Evaluation*, *21*, 121–130.

Mueller, M. R. (1998). The evaluation of Minnesota's Early Childhood Family Education Program: A dialogue. *American Journal of Evaluation*, *19*, 80–99.

Murray, F. B. (2009). An accreditation dilemma: The tension between program accountability and program improvement in programmatic accreditation. In P. M. O'Brian (Ed.), *Accreditation. Assuring and enhancing quality*. New Directions in Higher Education, No. 145, 59–68. San Francisco: Jossey-Bass.

National Commission on Excellence in Education (1983). *A nation at risk.* Washington, DC: U.S. Department of Education.

National Performance Review. (1993). *Reaching public goals: Managing government for results*. Washington, DC: Government Printing Office.

Nay, J., & Kay, P. (1982). *Government oversight and evaluability assessment.* Lexington, MA: Heath.

Neal, A. D. (October, 2008). Seeking higher-ed accountability: Ending federal accreditation. *Change. The Magazine of Higher Learning, 40*(5), 24–29.

Newcomer, K. E. (2001). Tracking and probing program performance: Fruitful path or blind alley

for evaluation professionals? *American Journal of Evaluation*, *22*, 337–342.

Nowakowski, J., Bunda, M. A., Working, R., Bernacki, G., & Harrington, P. (1985). *A handbook of educational variables*. Boston: Kluwer-Nijhoff.

Oakes, J. M. (2002). Risks and wrongs in social science research: An evaluator's guide to the IRB. *Evaluation Review*, *26*, 443–479.

O'Brien, P. M. (2009). Editor's notes. In P. M. O'Brien (Ed.), *Accreditation: Assuring and enhancing quality.* New Directions in Higher Education, No. 145, 1–6. San Francisco: Jossey-Bass.

Office of Management and Budget. (2004). *Program evaluation: What constitutes strong evidence of program effectiveness?* Retrieved March 13, 2009, from http://www.whitehouse.gov/omb/part/2004-program-eval.pdf.

OMB Watch. (2000). *GPRA: Background information*. Retrieved March 13, 2009, from http://www.ombwatch.org/node/326.

Osborne, D., & Gaebler, T. (1992). *Reinventing government: How the entrepreneurial spirit is transforming the public sector.* Reading, MA: Addison-Wesley.

O'Sullivan, E., Rassel, G. R., & Berner, M. (2003). *Research methods for public administrators* (4th ed.). White Plains, NY: Longman.

O'Sullivan, R. G., & D'Agostino, A. (2002). Promoting evaluation through collaboration: Findings from community-based programs for young children and their families. *Evaluation*, *8*(3), 372–387.

Owen, J. M., & Lambert, F. C. (1998). Evaluation and the information needs of organizational leaders. *American Journal of Evaluation*, *19*(3), 355–365.

Patton, M. Q. (1975). *Alternative evaluation research paradigm*. Grand Forks: North Dakota Study Group on Evaluation.

Patton, M. Q. (1986). *Utilization-focused evaluation* (2nd ed.). Beverly Hills, CA: Sage.

Patton, M. Q. (1988). Politics and evaluation. *Evaluation Practice*, *9*, 89–94.

Patton, M. Q. (1990). The challenge of being a profession. *Evaluation Practice*, *11*, 45–51.

Patton, M. Q. (1994). Developmental evaluation. *Evaluation Practice*, *15*(3), 311–319.

Patton, M. Q. (1996). A world larger than formative and summative. *Evaluation Practice*, *17*(2), 131–144.

Patton, M. Q. (1997a). Of vacuum cleaners and toolboxes: A response to Fetterman's response. *Evaluation Practice*, *18*(3), 267–270.

Patton, M. Q. (1997b). Toward distinguishing empowerment evaluation and placing it in a larger context. *Evaluation Practice*, *18*(2), 147–163.

Patton, M.Q. (1997c). *Utilization-focused evaluation: The new century text.* Thousand Oaks, CA: Sage.

Patton, M. Q. (2000). Overview: Language matters. In R. K. Hopson (Ed.), *How and why language matters in evaluation.* New Directions for Evaluation, No. 86, 5–16. San Francisco: Jossey-Bass.

Patton, M. Q. (2001). *Qualitative evaluation and research methods* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2005a). Toward distinguishing empowerment evaluation and placing it in a larger context: Take two. *American Journal of Evaluation*, *26*(3), 408–414.

Patton, M. Q. (2005b). Developmental evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.

Patton, M. Q. (2005c). Book review: *Empowerment evaluation principles in practice. American Journal of Evaluation, 26*(3), 408–414.

Patton, M. Q. (2005d). Patton responds to Fetterman, Wandersman, and Snell-Johns. *American Journal of Evaluation*, *26*(3), 429–430.

Patton, M. Q. (2008a). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2008b). sup wit eval ext? In M. T. Braverman, M. Engle, M. E. Arnold, & R. A. Rennekamp (Eds.), *Program evaluation in a complex organizational system: Lessons from Cooperative Extension.* New Directions for Evaluation, No. 120, 101–115. San Francisco: Jossey-Bass.

Patton, M. Q. (2009). Presidential strand session presented at the annual conference of the American Evaluation Association, Orlando, FL.

Patton, M. Q. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation and use.* New York: Guilford.

Patton, M. Q., King, J., & Greenseid, L. (2007). The oral history of evaluation Part V: An interview with Michael Quinn Patton. *American Journal of Evaluation*, *28*, 102–114.

Pawson, R. (2003). Nothing as practical as a good theory. *Evaluation*, *9*(3), 471–490.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Thousand Oaks, CA: Sage.

Perrin, B. (1998). Effective use and misuse of performance measurement. *American Journal of Evaluation*, *19*, 367–379.

Perry, J., Engbers, T., & Jun, S. (2009). Back to the future? Performance-related pay, empirical research, and the perils of persistence. *Public Administration Review*, *69*(1), 39–51.

Perry, K. M. (2008). A reaction to and mental metaevaluation of the Experiential Learning Evaluation Project. *American Journal of Evaluation*, *29*(3), 352–357.

Peters, G. B. (1999). *American public policy: Process and performance*. Chatham, NJ: Chatham House.

Posavac, E. J. (1994). Misusing program evaluation by asking the wrong question. In C. J. Stevens & M. Dial (Eds.), *Preventing the misuse of evaluation*. New Directions for Program Evaluation, No. 64, 69–78. San Francisco: Jossey-Bass.

Positer, T. H. (2004). Performance monitoring. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.). *Handbook of practical program evaluation* (2nd ed.). San Francisco: Jossey-Bass.

Preskill, H. (2008). Evaluation's second act: A spotlight on learning. *American Journal of Evaluation, 28*, 127–138.

Preskill, H., & Boyle, S. (2008). A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation*, *29*(4), 443–459.

Preskill, H., & Torres, R. T. (1998). *Evaluative inquiry for learning in organizations.* Thousand Oaks, CA: Sage.

Preskill, H., & Torres, R. T. (2000). The learning dimension of evaluation use. In V. J. Caracelli & H. Preskill (Eds.), *The expanding scope of evaluation use*. New Directions for Evaluation, No. 88. San Francisco: Jossey-Bass.

Provus, M. M. (1971). *Discrepancy evaluation.* Berkeley, CA: McCutchan.

Provus, M. M. (1973). Evaluation of ongoing programs in the public school system. In B. R. Worthen & J. R. Sanders (Eds.), *Educational evaluation: Theory and practice*. Belmont, CA: Wadsworth.

Radin, B. (2006). *Challenging the performance movement: Accountability, complexity, and democratic values*. Washington, DC: Georgetown University Press.

Rallis, S. F., & Rossman, G. B. (2000). Dialogue for learning: Evaluator as critical friend. In R. K. Hopson (Ed.), *How and why language matters in evaluation*. New Directions for Evaluation, No. 86, 81–92. San Francisco: Jossey-Bass.

Reason, P. (1994). Three approaches to participative inquiry. In N. K. Denzin and Y. S. Lincoln (Eds.), *Handbook of qualitative research.* Thousand Oaks: Sage.

Reichardt, C. S. (1994). Summative evaluation, formative evaluation, and tactical research. *Evaluation Practice*, *15*, 275–281.

Reichardt, C. S., & Rallis, S. F. (1994). Qualitative and quantitative inquiries are not incompatible: A call for a new partnership. In C. S. Reichardt & S. F. Rallis (Eds.), *The qualitative-quantitative debate: New perspectives*. New Directions for Program Evaluation, No. 61, 85–91. San Francisco: Jossey-Bass.

Reichardt, C. S., Trochim, W. K., & Cappelleri, J. C. (1995). Reports of the death of regression-discontinuity analysis are greatly exaggerated. *Evaluation Review*, *19*, 39–63.

Resnick, L. B. (2006). Making accountability really count. *Educational Measurement: Issues and Practice*, *25*(1), 33–37.

Rich, R. F., & Oh, C. H. (2000). Rationality and use of information in policy decisions. *Science Communication, 22,* 173–211.

Rist, R. C. (Ed.) (1990). *Program evaluation and the management of government: Patterns and prospects across eight nations.* New Brunswick, NJ: Transaction Publishers.

Rodosky, R. J., & Muñoz, M. A. (2009). Slaying myths, eliminating excuses: Managing for accountability by putting kids first. In D. W. Compton & M. Braizerman (Eds.), *Managing program evaluation: Towards explicating a professional practice.* New Directions for Evaluation, No. 121, 43–54. San Francisco: Jossey-Bass.

Rodriguez-Campos, L. (2005). *Collaborative evaluations: A step-by-step model for the evaluator*. Tarmac, FL: Lumina Press.

Rog, D. J. (1985). *A methodological analysis of evaluability assessment*. Unpublished doctoral dissertation. Vanderbilt University.

Rog, D. J. (1994). *The homeless families program: Interim benchmarks*. Washington, DC: Vanderbilt Institute for Public Policy Studies.

Rog, D. J., Holupka, C. S., McCombs-Thornton, K., Brito, M. C., & Hambrick, R. (1997). Case management in practice: Lessons from the evaluation of RWJ/HUD Homeless Families Program. *Journal of Prevention and Intervention in the Community*, *15*, 67–82.

Rogers, E. M. (1995). *Diffusion of innovations* (4th ed.). New York: Free Press.

Rogers, P. J. (2000). Program theory: Not whether programs work but how they work. In D. Stufflebeam, G. Madaus, & T. Kellaghan (Eds.), *Evaluation models.* Boston: Kluwer Academic.

Rogers, P. J. (2001). The whole world is evaluating half-full glasses. *American Journal of Evaluation, 22,* 431–436.

Rogers, P. J. (2007). Theory-based evaluation: Reflections ten years on. In S. Mathison (Ed.), *Enduring issues in evaluation: The 20th anniversary of the collaboration between NDE and AEA.* New Directions for Evaluation, No. 114, 63–66. San Francisco: Jossey-Bass.

Rogers, P. J., & Williams, B. (2006). Evaluation for practice improvement and organizational learning. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation.* Thousand Oaks, CA: Sage.

Rosenbaum, D. P., & Hanson, G. S. (1998). Assessing the effects of school-based drug education: A six-year multi-level analysis of project D.A.R.E. *Journal of Crime and Delinquency*, *35*, 381–412.

Rossi, P. H. (1971). Boobytraps and pitfalls in the evaluation of social action programs. In F. G. Caro (Ed.), *Readings in evaluation research*. New York: Sage.

Rossi, P. H., & Freeman, H. E. (1985). *Evaluation: A systematic approach* (3rd ed.). Beverly Hills, CA: Sage.

Rossi, P. H., Freeman, H. E., & Lipsey, M. E. (1998). *Evaluation: A systematic approach* (6th ed.). Newbury Park: CA: Sage.

Rossi, P. H., Lipsey, M. E., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Rouge, J. C. (2004). The origin and development of the African Evaluation Guidelines. In C. Russon & G. Russon (Eds.), *International perspectives on evaluation standards.* New Directions for Evaluation, No. 104, 55–66. San Francisco: Jossey-Bass.

Rubin, H. J., & Rubin, I. S. (1995). *Qualitative interviewing.* Thousand Oaks, CA: Sage.

Russ-Eft, D. F., & Preskill, H. S. (2001). *Evaluation in organizations: A systematic approach to enhancing learning, performance, and change*. Cambridge, MA: Perseus.

Ryan, K. E., & DeStefano, L. (Eds.). (2000a). *Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation.* New Directions for Evaluation, No. 85, San Francisco: Jossey-Bass.

Ryan, K. E., & DeStefano, L. (2000b). Disentangling dialogue. Lessons from practice. In K. E. Ryan & L. DeStefano (Eds.), *Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation.* New Directions for Evaluation, No. 85, 63–76. San Francisco: Jossey-Bass.

Ryan, K. E., & Johnson, T. D. (2000). Democratizing evaluation: Meaning and methods from practice. In K. E. Ryan & L. DeStefano (Eds.), *Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation*. New Directions for Evaluation, No. 85, 39–50. San Francisco: Jossey-Bass.

Sackett, D. L., & Wennberg, J. E. (1997). Choosing the best research design for each question (editorial). *British Medical Journal*, *315*, 1636.

Salkind, N. J. (2010). *Excel statistics: A quick guide.* Thousand Oaks, CA: Sage.

Sanders, J. (2002). Presidential address: On mainstreaming evaluation. *American Journal of Evaluation*, *23*(3), 253–259.

Sanders, J. R. (1995). Standards and principles. In W. R. Shadish, D. L. Newman, M. A. Scheirer, & C. Wye (Eds.) *Guiding principles for evaluators.* New Directions for Program Evaluation, No. 66, 47–52. San Francisco: Jossey-Bass.

Sanders, J. R. (2001). A vision for evaluation. *American Journal of Evaluation*, *22*, 363–366.

Sanders, J., & Sullins, C. (2005). *Evaluating school programs* (3rd ed.). Thousand Oaks, CA: Corwin.

Schnoes, C. J., Murphy-Berman, V., & Chambers, J. M. (2000). Empowerment evaluation applied: Experiences, analysis and recommendations from a case study. *American Journal of Evaluation*, *21*, 53–64.

Schwandt, T. A. (1997). Reading the "problem of evaluation" in social inquiry. *Qualitative Inquiry*, *3*(1), 4–25.

Schwandt, T. A. (1999). *Dialogue in evaluation: Philosophy, theory, and practice*. Presentation at the annual meeting of the American Evaluation Association, Orlando, FL.

Schwandt, T. A. (2001a). Responsiveness and everyday life. In J. C. Greene & T. A. Abma (Eds.), *Responsive evaluation*. New Directions for Evaluation, No. 92, 73–88. San Francisco: Jossey-Bass.

Schwandt, T. A. (2001b). *Dictionary of qualitative inquiry* (2nd ed.). Thousand Oaks, CA: Sage.

Schwandt, T. A. (2004). *On methods and judgments in evaluation.* Plenary presentation at the meeting of the American Evaluation Association, Atlanta, Ga.

Schwandt, T. A. (2005). A diagnostic reading of scientifically-based research for education. *Educational Theory*, *55*, 285–305.

Schwandt, T. A. (2007). Thoughts on using the notion of evidence in the controversy over methods choice. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation methodology: Building the evidence base for method choice in government sponsored evaluation*. New Directions for Evaluation, No. 113, 115–119. San Francisco: Jossey-Bass.

Schwandt, T. A. (2008). Educating for intelligent belief in evaluation. *American Journal of Evaluation*, *29*(2), 139–150.

Schwandt, T. A., & Halpern, E. S. (1989). *Linking auditing and metaevaluation.* Newbury Park, CA: Sage.

Schwartz, S., & Baum, S. (1992). Education. In C. T. Clotfelder (Ed.), *Who benefits from the nonprofit sector?* Chicago: University of Chicago Press.

Scott, A., & Sechrest, L. (1992). Theory-driven approaches to benefit cost analysis: Implications of program theory. In H. Chen & P. H. Rossi (Eds.), *Using theory to improve program and policy evaluations*. New York: Greenwood.

Scriven, M. (1967). The methodology of evaluation. In R. E. Stake (Ed.), *Curriculum evaluation.* (American Educational Research Association Monograph Series on Evaluation, No. 1, pp. 39–83). Chicago: Rand McNally.

Scriven, M. (1972). Pros and cons about goal-free evaluation. *Evaluation Comment*, *3*, 1–7.

Scriven, M. (1973). The methodology of evaluation. In B. R. Worthen & J. R. Sanders (Eds.), *Educational evaluation: Theory and practice.* Belmont, CA: Wadsworth.

Scriven, M. (1974a). Evaluation perspectives and procedures. In W. J. Popham (Ed.), *Evaluation in education*. Berkeley, CA: McCutchan.

Scriven, M. (1974b). Standards for the evaluation of educational programs and products. In G. D. Borich (Ed.), *Evaluating educational programs and products*. Englewood Cliffs, NJ: Educational Technology.

Scriven, M. (1975). *Evaluation bias and its control.* Berkeley: University of California. Retrieved February 17, 2009, from http://www.wmich.edu/evalctr/pubs/ops/ops04.html.

Scriven, M. (1980). *The logic of evaluation*. Interness, CA: Edgepress.

Scriven, M. (1983). Evaluation ideologies. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation.* Boston: Kluwer-Nijhoff.

Scriven, M. (1991a). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.

Scriven, M. (1991b). Beyond formative and summative evaluation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century*. Ninetieth Yearbook of the National Society for the Study of Education (pp. 19–64). Chicago: University of Chicago Press.

Scriven, M. (1991c). Key evaluation checklist. In M. Scriven (Ed.), *Evaluation thesaurus.* Thousand Oaks, CA: Sage.

Scriven, M. (1993). *Hard-won lessons in program evaluation*. New Directions for Program Evaluation, No. 58, 1–107. San Francisco: Jossey-Bass.

Scriven, M. (1996). Types of evaluation and types of evaluator. *Evaluation Practice*, *17*, 151–162.

Scriven, M. (2005a). Book review, *Empowerment Evaluation Principles in Practice. American Journal of Evaluation, 26*(3), 415–417.

Scriven, M. (2005b). A note on David Fetterman's response. *American Journal of Evaluation, 26*(3), 431.

Scriven, M. (2007). *Key evaluation checklist.* http:// www.wmich.edu/evalctr/checklists/kec-feb07.pdf.

Scriven, M., Miller, R., & Davidson, J. (2005). The oral history of evaluation Part III: The professional evolution of Michael Scriven. *American Journal of Evaluation*, *26*, 378–388.

Sechrest, L. (1997). Review of the book *Empowerment Evaluation Knowledge and Tools for Self-Assessment and Accountability*. *Environment and Behavior*, *29*(3), 422–426.

Sechrest, L., & Figueredo, A. J. (1993). Program evaluation. *Annual Review of Psychology*, *44*, 645–674.

Sedlak, A. J., Mettenburg, J., Basena, M., Petta, I., McPherson, K., Greene, A., & Li, S. (2010). *Fourth national incidence study of child abuse and neglect* (NIS-4): Report to Congress, Executive Summary. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

Senge, P. M. (1990). *The fifth discipline: The art and practice of organizational learning.* New York: Doubleday.

SenGupta, S., Hopson, R., & Thompson-Robinson, M. (2004). Cultural competence in evaluation: An overview. In M. Thompson-Robinson, R. Hopson, & S. SenGupta (Eds.), *In search of cultural competence in evaluation: Toward principles and practices.* New Directions for Evaluation, No. 102, 5–20. San Francisco: Jossey-Bass.

Shadish, W. R. (1984). Lessons from the implementation of deinstitutionalization. *American Psychologist*, *39*, 725–738.

Shadish, W. R. (1994). Need-based evaluation theory: What do you need to know to do good evaluation? *Evaluation Practice*, *15*, 347–358.

Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation, 19*(1), 1–19.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation*. Newbury Park, CA: Sage.

Shadish, W., & Miller, R. (2003). The oral history of evaluation Part I: Reflections on the chance to work with great people: An interview with William Shadish. *American Journal of Evaluation, 24*(2), 261–272.

Shadish, W. R., Newman, D. L., Scheirer, M. A., & Wye, C. (Eds.). (1995). *Guiding principles for evaluators*. New Directions for Program Evaluation, No. 66. San Francisco: Jossey-Bass.

Shepard, L. A., & Peressini, D. D. (2002, January). An analysis of the content and difficulty of the CSAP 10th-grade Mathematics Test: A report to the Denver Area School Superintendents' Council (DASSC). Boulder, CO: School of Education, University of Colorado at Boulder. Available at

http://www.colorado.edu/education/faculty/lorrieshepard/reforms.html

Sieber, J. E. (1980). Being ethical: Professional and personal decisions in program evaluation. In R. E. Perloff & E. Perloff (Eds.), *Values, ethics, and standards in evaluation*. New Directions for Program Evaluation, No. 7, 51–61. San Francisco: Jossey-Bass.

Smith, E. R., & Tyler, R. W. (1942*). Appraising and recording student progress*. New York: Harper & Row.

Smith, M. F. (1989). *Evaluability assessment: A practical approach*. Boston: Kluwer Academic.

Smith, M. F. (2001). Evaluation: Preview of the future #2. *American Journal of Evaluation*, *22*, 281–300.

Smith, M. F. (2005). Evaluability assessment. In S. Mathison (Ed.), *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.

Smith, N. L. (1983). *Dimensions of moral and ethical problems in evaluation* (Paper and Report Series No. 92). Portland, OR: Northwest Regional Educational Laboratory, Research on Evaluation Program.

Smith, N. L. (1997, November). *An investigative framework for characterizing evaluation practice.* In N. L. Smith (Chair), Examining evaluation practice. Symposium conducted at the meeting of the American Evaluation Association, San Diego, CA.

Smith, N. L. (1998). Professional reasons for declining an evaluation contract. *American Journal of Evaluation*, *19*, 177–190.

Smith, N. L. (2004). Evidence and ideology. Paper presented at the meeting of the Canadian Evaluation Society, Saskatoon, Canada.

Smith, N. L. (2007). Empowerment evaluation as evaluation ideology. *American Journal of Evaluation*, *28* (2), 169–178.

Society for Prevention Research. (2004). *Standards of evidence: Criteria for efficacy, effectiveness, and dissemination.* Retrieved January 29, 2010, from http://www.preventionresearch.org/commlmon.php#SofE.

Sonnichsen, R. C. (1987). An internal evaluator responds to Ernest House's views on internal evaluation. *American Journal of Evaluation, 8*(4), 34–36.

Sonnichsen, R. C. (1999). *High impact internal evaluation.* Thousand Oaks, CA: Sage.

Spiegel, A. N., Bruning, R. H., & Giddings, L. (1999). Using responsive evaluation to evaluate a professional conference. *American Journal of Evaluation*, *20*, 57–67.

Springer, M. G., & Winters, M. A. (2009). *New York City's Bonus Pay Program: Early evidence from a randomized trial.* Vanderbilt: National Center on Performance Incentives.

St. Pierre, T. L., & Kaltreider, L. (2004). Tales of refusal, adoption, and maintenance: Evidence-based substance abuse prevention via school-extension collaborations. *American Journal of Evaluation*, *25*(4), 479–491.

Stake, R. E. (1967). The countenance of educational evaluation. *Teachers College Record*, *68*, 523–540.

Stake, R. E. (1969). Evaluation design, instrumentation, data collection, and analysis of data. In J. L. Davis (Ed.), *Educational evaluation*. Columbus, OH: State Superintendent of Public Instruction.

Stake, R. E. (1970). Objectives, priorities, and other judgment data. *Review of Educational Research*, *40*, 181–212.

Stake, R. E. (1973). Program evaluation, particularly responsive evaluation. Keynote address at the conference "New Trends in Evaluation," Institute of Education, University of Goteborg, Sweden, Oct. 1973. In G. F. Madaus, M. S. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation.* Boston: Kluwer-Nijhoff, 1987.

Stake, R. E. (1975a). *Evaluating the arts in education: A responsive approach.* Columbus, OH: Merrill.

Stake, R. E. (1975b). Program evaluation, particularly responsive evaluation (Occasional Paper No. 5). Kalamazoo: Western Michigan University Evaluation Center.

Stake, R. E. (1978). The case study method in social inquiry. *Educational Researcher*, *7*, 5–8.

Stake, R. E. (1980). Program evaluation, particularly responsive evaluation. In W. B. Dockrell & D. Hamilton (Eds.), *Rethinking educational research.* London: Hodeder & Stoughton.

Stake, R. E. (1991). Retrospective on "The countenance of educational evaluation." In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century.* Ninetieth Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press.

Stake, R. E. (1994). Case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research.* Thousand Oaks, CA: Sage.

Stake, R. E. (1995). *The art of case study research.* Thousand Oaks, CA: Sage.

Stake, R. E. (2000a). Case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Stake, R. E. (2000b). A modest commitment to the promotion of democracy. In K. E. Ryan & L. DeStefano (Eds.), *Evaluation as a democratic process: Promoting inclusion, dialogue, and*

deliberation. New Directions for Evaluation, No. 85, 97–106. San Francisco: Jossey-Bass.

Stake, R. E., & Schwandt, T. A. (2006). On discerning quality in evaluation. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation*. Thousand Oaks, CA: Sage.

Standaert, R. (2000). Inspectorates of education in Europe: A critical analysis. Utrecht: Standing Interim Conference of Central and General Inspectorates of Education. Retrieved March 8, 2009, from http://www.sici-inspectorates.org/shared/data/pdf/inspectorates-20of-20education-20in-20-europe-jan00.pd.

Stephan, A. S. (1935). *Prospects and possibilities: The New Deal and the new social research*. Chapel Hill: University of North Carolina Press.

Stevahn, L., King, J. A., Ghere, G., & Minnema, J. (2005). Establishing essential competencies for program evaluators. *American Journal of Evaluation*, *26*(1), 43–59.

Stevenson, J., & Thomas, D. (2006). Intellectual contexts. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation.* Thousand Oaks, CA: Sage.

Stockdill, S. H., Braizerman, M., & Compton, D. W. (2002). Toward a definition of the ECB process: A conversation with the ECB literature. In D. W. Compton, M. Braizerman, & S. H. Stockdill (Eds.), *The art, craft, and science of evaluation capacity building.* New Directions for Evaluation, No. 93, 7–26. San Francisco: Jossey-Bass.

Strauss, A., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage.

Stufflebeam, D. L. (1968). *Evaluation as enlightenment for decision making.* Columbus: Ohio State University Evaluation Center.

Stufflebeam, D. L. (1971). The relevance of the CIPP evaluation model for educational accountability. *Journal of Research and Development in Education*, *5*, 19–25.

Stufflebeam, D. L. (1973a). Excerpts from "Evaluation as enlightenment for decision making." In B. R. Worthen & J. R. Sanders (Eds.), *Educational evaluation: Theory and practice.* Belmont, CA: Wadsworth.

Stufflebeam, D. L. (1973b). An introduction to the PDK book: Educational evaluation and decision-making. In B. R. Worthen & J. R. Sanders (Eds.), *Educational evaluation: Theory and practice*. Belmont, CA: Wadsworth.

Stufflebeam, D. L. (1974). *Metaevaluation* (Occasional Paper No. 3). Kalamazoo: Western Michigan University Evaluation Center.

Stufflebeam, D. L. (1977, April). Working paper on needs assessment in evaluation. Paper presented at the American Educational Research Association Evaluation Conference, San Francisco.

Stufflebeam, D. L. (1994). Empowerment evaluation, objectivist evaluation, and evaluation standards: Where the future of evaluation should not go and where it needs to go. *Evaluation Practice*, *15*, 321–338.

Stufflebeam, D. L. (1999). Evaluation contracts checklist. Retrieved February 15, 2010, from http://www.wmich.edu/evalctr/checklists/contracts.pdf.

Stufflebeam, D. L. (2000). Lessons in contracting for evaluations. *American Journal of Evaluation*, *21*, 293–314.

Stufflebeam, D. L. (2001a). Evaluation checklists: Practical tools for guiding and judging evaluations. *American Journal of Evaluation*, *22*, 71–79.

Stufflebeam, D. L. (2001b). *Evaluation models.* New Directions for Evaluation, No. 89. San Francisco: Jossey-Bass.

Stufflebeam, D. L. (2001c). The metaevaluation imperative. *American Journal of Evaluation*, *22*(2), 183–209.

Stufflebeam, D. L. (2002a). *Institutionalizing evaluation checklist.* Retrieved February 17, 2008, from http://www.wmich.edu/evalctr/checklists/institutionalizingeval.pdf.

Stufflebeam, D. L. (2002b). CIPP evaluation model checklist. Retrieved from http://www.wmich.edu/evalctr/checklists.

Stufflebeam, D. L. (2004a). A note on the purposes, development and applicability of the Joint Committee Evaluation Standards. *American Journal of Evaluation*, *25*, 99–102.

Stufflebeam, D. L. (2004b). The 21st-century CIPP model. In M. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Stufflebeam, D. L. (2005). CIPP model (context, input, process, product). In S. Mathison (Ed.), *Encyclopedia of evaluation.* Thousand Oaks, CA: Sage.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus, M. M. (1971). *Educational evaluation and decision making*. Itasca, IL: Peacock.

Stufflebeam, D. L., Madaus, G., & Kellaghan, T. (Eds.). (2000). *Evaluation models*. Boston: Kluwer Academic.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications*. San Francisco: Jossey-Bass.

Stufflebeam, D. L., & Wingate, L. A. (2005). A self-assessment procedure for use in evaluation training. *American Journal of Evaluation*, *26*, 544–561.

Suchman, E. (1967). *Evaluative research.* New York: Sage.

Sylvia, R. D., Meier, K. J., & Gunn, E. M. (1985). *Program planning and evaluation for the public manager.* Monterey, CA: Brooks/Cole.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Neeedham Heights, MA: Allyn & Bacon.

Talmage, H. (1982). Evaluation of programs. In H. E. Mitzel (Ed.), *Encyclopedia of educational research* (5th ed.). New York: Free Press.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches.* Applied Social Research Methods, No. 46. Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and behavioral research.* Thousand Oaks, CA: Sage.

Taylor-Powell, E. & Boyd, H. H. (2008). Evaluation capacity building in complex organizations. In M. T. Braverman, M. Engle, M. E. Arnold, & R. A. Rennekamp (Eds.), *Program evaluation in a complex organizational system: Lessons from cooperative extension*. New Directions for Evaluation, No. 120, 55–69. San Francisco: Jossey-Bass.

Teske, P., Fitzpatrick, J. L., & Kaplan, G. (2006). The information gap? *Review of Policy Research*, *23*, 969–981.

Tharp, R. G., & Gallimore, R. (1979). The ecology of program research and evaluation: A model of evaluation succession. In L. Sechrest, S. G. West, M. A. Phillips, R. Rechner, & W. Yeaton (Eds.), *Evaluation Studies Review Annual*, *4*, 39–60.

Thompson-Robinson, M., Hopson, R., & SenGupta, S. (Eds.). (2004). *In search of cultural competence in evaluation: Toward principles and practices.* New Directions for Evaluation, No. 102. San Francisco: Jossey-Bass.

Torres, R. T., Padilla Stone, S., Butkus, D. L., Hook, B. B., Casey, J., & Arens, S. A. (2000). Dialogue and reflection in a collaborative evaluation: Stakeholder and evaluator voices. In K. E. Ryan & L. DeStefano (Eds.), *Evaluation as a democratic process: Promoting inclusion, dialogue, and deliberation.* New Directions for Evaluation, No. 85, 27–38. San Francisco: Jossey-Bass.

Torres, R. T., Preskill, H., & Piontek, M. (2005). *Evaluation strategies for communicating and reporting: Enhancing learning in organizations* (2nd ed.). Newbury Park, CA: Sage.

Toulemond, J. (2009). *Personal communication.*

Trochim, W. M. K. (1984*). Research design for program evaluation: The regression-discontinuity approach*. Newbury Park, CA: Sage.

Trochim, W. M. K., & Linton, R. (1986). Conceptualization for planning and evaluation. *Evaluation and Program Planning*, *9*, 289–308.

Trochim, W. M. K., Mark, M. M., & Cooksy, L. J. (Eds.) (2009). *Evaluation policy and evaluation practice.* New Directions for Evaluation, No. 123. San Francisco: Jossey-Bass.

Tyler, R. W. (1942). General statement on evaluation. *Journal of Educational Research*, *35*, 492–501.

Tyler, R. W. (1950). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.

Tyler, R. W. (1991). General statement on program evaluation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century.* Ninetieth Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press.

United Way of America. (1996). *Measuring program outcomes: A practical approach*. Alexandria, VA: United Way of America.

U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education.* Washington, DC. Retrived June 1, 2010, from http://www2.ed.gov/about/bdscomm/list/hiedfuture/reports/final-report.pdf.

Van Mondfrans, A. (1985). *Guidelines for reporting evaluation findings.* Unpublished manuscript, Brigham Young University, College of Education, Provo, UT.

Vestman, O. K., & Conners, R. F. (2006). The relationship between evaluation and politics. In I. F. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The Sage handbook of evaluation*. Thousand Oaks, CA: Sage.

Vroom, P. E., Colombo, M., & Nahan, N. (1994). Confronting ideology and self-interest: Avoiding misuse of evaluation. In C. J. Stevens & M. Dial (Eds.), *Preventing the misuse of evaluation*. New Directions for Program Evaluation, No. 64, 49–59. San Francisco: Jossey-Bass.

Wadsworth, Y. (2001). Becoming responsive—and some consequences for evaluation as dialogue across distance. In J. C. Greene & T. A. Abma (Eds.), *Responsive evaluation.* New Directions for Evaluation, No. 92, 45–58. San Francisco: Jossey-Bass.

Wallis, A., Dukay, V., & Fitzpatrick, J. (2008). Evaluation of Godfrey's Children Center in Tanzania. In J. Fitzpatrick, C. Christie, & M. Mark (Eds.), *Evaluation in action: Interviews with expert evaluators*. Thousand Oaks, CA: Sage.

Wandersman, A., Flaspohler, P., Ace, A., Ford, L., Imm, P. S., Chinman, M. J., Sheldon, J., Bowers Andrews, A., Crusto, C. A., & Kaufman, J. S. (2003). PIE à la mode: Mainstreaming evaluation and accountability in each program in every county of a statewide school readiness initiative. In J. J. Barnette, & J. R. Sanders (Eds.), *The mainstreaming of evaluation.* New Directions for Evaluation, No. 99, 33–50. San Francisco: Jossey-Bass.

Wandersman, A., Imm, P.S., Chinman, M., & Kafterian, S. (2000). Getting To Outcomes: A results-based approach to accountability. *Evaluation and Program Planning, 23*, 389–395.

Wandersman, A., & Snell-Johns, J. (2005). Empowerment evaluation: Clarity, dialogue, and growth. *American Journal of Evaluation, 26*(3), 421–428.

Weaver, L., & Cousins, J. B. (2004). Unpacking the participatory process. *Journal of Multidisciplinary Evaluation, 1*, 19–40.

Weiss, C. H. (1972). *Evaluation research: Methods for assessing program effectiveness.* Englewood Cliffs, NJ: Prentice-Hall.

Weiss, C. H. (1973). Where politics and evaluation research meet. *Evaluation, 1*, 37–45.

Weiss, C. H. (1977). *Using social research in public policy making.* Lexington, MA: Lexington Books.

Weiss, C. H. (1983). Ideology, interests, and information. In D. Callahan & B. Jennings (Eds.), *Ethics, the social sciences, and policy analysis* (pp. 231–245). New York: Plenum Press.

Weiss, C. H. (1987). Evaluating social programs: What have we learned? *Society, 25*, 40–45.

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. P. Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), *New approaches to evaluating community initiatives Volume 1: Concepts, methods, and context*. (pp. 65–92). Washington, DC: Aspen Institute.

Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. In D. Rog & D. Fournier (Eds.), *Progress and future directions in evaluation: Perspectives on theory, practice, and methods*. New Directions for Evaluation, No. 76. San Francisco: Jossey-Bass.

Weiss, C. H. (1998a). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Weiss, C. H. (1998b). Have we learned anything new about the use of evaluation? *American Journal of Evaluation, 19*, 21–33.

Weiss, C. H., & Bucuvalas, M. J. (1980a). Truth tests and utility tests: Decision-makers' frames of reference for social science research. *American Sociological Review, 45*, 302–313.

Weiss, C. H., & Bucuvalas, M. J. (1980b). *Social science research and decision-making*. New York: Columbia University Press.

Weiss, C. H., & Mark, M. M. (2006). The oral history of evaluation Part IV: The professional evolution of Carol Weiss. *American Journal of Evaluation, 27*(4), 475–483.

Wholey, J. S. (1983). *Evaluation and effective public management*. Boston: Little, Brown & Co.

Wholey, J. S. (1986). Using evaluation to improve government performance. *Evaluation Practice, 7*, 5–13.

Wholey, J. S. (1994). Assessing the feasibility and likely usefulness of evaluation. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation*. San Francisco: Jossey-Bass.

Wholey, J. S. (1996). Formative and summative evaluation: Related issues in performance measurement. *Evaluation Practice, 17*, 145–149.

Wholey, J. S. (1999a). Performance-based management: Responding to the challenges. *Public Productivity and Management Review, 22*(3), 288–307.

Wholey, J. S. (1999b). Quality control: Assessing the accuracy and usefulness of performance measurement systems. In H. P. Hatry (Ed.), *Performance measurement: Getting results* (pp. 217–237). Washington, DC: Urban Institute.

Wholey, J. S. (2001). Managing for results: Roles for evaluators in a new management era. *American Journal of Evaluation, 22*(4), 343–347.

Wholey, J. S. (2003). Improving performance and accountability: Responding to emerging management challenges. In S. J. Donaldson & M. Scriven (Eds.), *Evaluating social programs and problems* (pp. 43–61). Mahwah, NJ: Lawrence Erlbaum.

Wholey, J. S. (2004a). Using evaluation to improve performance and support policy decision making. In M. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Wholey, J. S. (2004b). Evaluability assessment. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (2nd ed.). San Francisco: Jossey-Bass.

Wholey, J. S., Hatry, H. P., & Newcomer, K. E. (Eds.). (2004). *Handbook of practical program evaluation* (2nd ed.). San Francisco: Jossey-Bass.

Wholey, J. S., & White, B. F. (1973). Evaluation's impact on Title I elementary and secondary education program management. *Evaluation, 1*, 73–76.

Williams, A., & Giardina, E. (Eds.). (1993). *The theory and practice of cost-benefit analysis*. Brookfield, VT: Edward Elgar.

Winston, J. (1999). Understanding performance measurement: A response to Perrin. *American Journal of Evaluation*, *20*, 95–100.

Witkin, B. R., & Altschuld, J. W. (1995). *Planning and conducting needs assessments.* Thousand Oaks, CA: Sage.

Worthen, B. R. (1975). Competencies for educational research and evaluation. *Educational Researcher*, *4*, 13–16.

Worthen, B. R. (1977, April). Eclecticism and evaluation models: Snapshots of an elephant's anatomy? Paper presented at the annual meeting of the American Educational Research Association, New York.

Worthen, B. R. (1995). Some observations about the institutionalization of evaluation. *Evaluation Practice*, *16*, 29–36.

Worthen, B. R. (1996). A survey of Evaluation Practice readers. *Evaluation Practice*, *17*, 85–90.

Worthen, B. R. (2001). Whither evaluation? That all depends. *American Journal of Evaluation*, *22*, 409–498.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). *Measurement and evaluation in the schools.* White Plains, NY: Longman.

Worthen, B. R., & Sanders, J. R. (1973). *Educational evaluation: Theory and practice.* Belmont, CA: Wadsworth.

Yates, B. T. (1996). *Analyzing costs, procedures, processes, and outcomes in human services.* Thousand Oaks, CA: Sage.

Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Thousand Oaks, CA: Sage.

Yin, R. K., & Davis, D. (2007). Adding new dimensions to case study evaluations: The case of evaluating comprehensive reforms. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation methodology: Building the evidence base for method choice in government sponsored evaluation.* New Directions for Evaluation, No. 113, 75–94. San Francisco: Jossey-Bass.

Yin, R. K., Davis, D., & Schmidt, R. J. (2005). The cross-site evaluation of the urban systemic program. Final report, strategies and trends in urban education reform. Bethesda, MD: COSMOS Corporation.

Yin, R. K., & Kaftarian, S. J. (1997). Introduction: Challenges of community-based program outcome evaluations. *Evaluation and Program Planning, 20*(3), 293–297.

Zvoch, K. (2009). Treatment fidelity in multisite evaluation: A multilevel longitudinal examination of provider adherence status and change. *American Journal of Evaluation*, *30*(1), 44–61.

Author Index

Abma, T., 192, 196
Abma, T. A., 321
Ackerman, F., 180
Agar, M., 436
Alkin, M. C., 9, 109, 122, 139, 143, 144, 173, 177, 178, 179, 182, 206, 244, 473
Altschuld, J. W., 263
American Evaluation Association, 15, 166, 238, 240, 335
Amie, M., 82
Andersen, L. A., 240
Anisur-Rahman, M., 201
Arens, S. A., 219, 222
Argyris, C., 205
Avery, M. E., 237

Babbie, E., 445
Baizerman, M., 237, 239
Baker, A., 14
Bandura, A., 205
Barbour, R., 439
Barela, E., 29, 236, 273, 285
Barnett, W. S., 413
Barnette, J. J., 236
Bell, J. B., 357, 363
Bell, W., 45
Benjamin, L. M., 485
Benkofske, M., 467
Bernard, H. R., 434, 446
Berner, M., 394, 407
Bickman, L., 17, 107, 160, 161, 162, 171, 225, 262, 284, 292, 416
Bledsoe, K., 18, 107, 171, 214, 224, 242, 299, 313, 489
Bloom, B. S., 42, 165
Bloom, G. A., 200, 202, 220, 247
Borg, W. R., 348
Boruch, R., 400
Botcheva, L., 233
Boyd, H. H., 160, 276, 292, 293, 294
Boyle, S., 60, 238, 239
Brandon, P. R., 317, 318, 426
Braverman, M. T., 427, 434
Brinkerhoff, R. O., 355, 391, 461
Brito, M. C., 16
Brittingham, B., 131, 132, 133
Bruner, B., 14
Bruning, R. H., 344
Buchanan, G. N., 268
Buckley, J., 56
Bucuvalas, M. J., 205, 224, 454, 462, 481
Butkus, D. L., 222

Calhoun, E. F., 12
Campbell, D. T., 46, 115, 267, 343, 404, 476
Campbell, R., 212, 213
Cappelleri, J. C., 404
Caracelli, V. J., 386
Carman, J. G., 51
Caro, F. G., 46
Carr, W., 203
Casey, J., 222
Casey, M., 438, 439
Chambers, J. M., 325
Chapel, T. J., 237
Chelimsky, E., 15, 17, 29, 66, 69, 71, 72, 73, 74, 244, 383, 396, 455
Chen, H. T., 17, 26, 160, 161, 162, 276, 323
Chinman, M., 211
Christian, L. M., 430
Christie, C., 215, 230
Christie, C. A., 29, 37, 52, 112, 122, 139, 177, 178, 220, 236, 244, 247, 273, 285, 306, 313, 376
Chubb, J. E., 56
Cohn, R., 79, 80, 81, 88, 94, 95
Colombo, M., 290, 300
Compton, D. W., 237, 239
Conner, R. F., 83, 215, 230, 306, 376
Conners, R. F., 69, 70
Converse, P. D., 433
Cook, D. L., 362
Cook, T. D., 45, 115, 120, 143, 144, 196, 343, 404, 480, 486
Cooksey, L. J., 94, 98, 493
Corbin, J. M., 446
Cotton, D. A., 237
Cousins, J. B., 180, 181, 199, 200, 201, 202, 203, 205, 206, 209, 212, 214, 220, 221, 222, 224, 247, 317, 321, 332, 462, 480, 481, 485
Covert, R. W., 277
Cox, J., 430
Cox, P. L., 317
Crary, D., 389
Creswell, J. W., 385
Cronbach, L. J., 7, 39, 40, 41, 43, 66, 262, 263, 304, 315, 328, 329, 336
Cullen, A., 199, 202, 212, 221, 224, 247

D'Agostino, A., 199
Dahler-Larsen, P., 70, 104, 165, 167, 226, 236, 240
Datta, L., 51, 52, 70, 236, 396, 492
Datta, L. E., 66
Davidson, J., 332, 346
Davis, D., 406
Davis, D. E., 58
deLeon, L., 57
Denhardt, R., 57

Denzin, N. K., 117, 437
DeStefano, L., 218, 219
Dewey, J. D., 71
Dickeson, R. C., 133
Dillman, D. A., 430, 431
Dodson, S. C., 158
Donaldson, S. I., 15, 17, 23, 49, 77, 110, 153, 160, 162, 163, 167, 171, 300, 313, 323, 489, 492
Donmoyer, R., 142
Donohue, J. J., 200, 202, 220, 247
Duignan, P., 236, 237
Dunet, D. O., 370

Earl, L. M., 180, 181, 199, 201, 203, 205, 206
Eden, C., 180
Eisner, E. W., 139, 140, 141
Ellett, F., 122
Elliott, J., 13
Engbers, T., 118, 324
Engelhart, M. D., 42

Fals-Borda, O., 201
Fetterman, D., 157, 188, 225, 245, 285, 300, 307, 335, 417, 424, 460
Fetterman, D. M., 201, 209, 210, 211, 212, 214, 215, 222, 350, 460
Fidell, L. S., 343
Figueredo, A. J., 65, 118
Fink, A., 430
Fischer, C. T., 473
Fitzpatrick, J. L., 12, 18, 23, 32, 37, 56, 77, 82, 89, 93, 107, 137, 152, 157, 167, 171, 188, 214, 222, 224, 225, 242, 244, 245, 247, 262, 284, 285, 299, 300, 302, 307, 313, 323, 338, 339, 352, 416, 417, 424, 460, 488, 489, 495
Flexner, A., 40, 129, 130
Floden, R. E., 129
Fontana, A., 438
Frankfort-Nachmias, C., 115
Frechtling, J., 233
Fredericks, K. A., 51
Freeman, H. E., 39, 40
Freire, P., 117, 201
Frey, J. H., 438
Fricker, R. D., 433
Frierson, H., 233
Furst, E. J., 42

Gaebler, T., 57
Gallimore, R., 267
Geiger, F. L., 473
Gephart, W. J., 111
Ghere, G., 12, 277
Giardina, E., 414
Giddings, L., 344
Gilmour, J. B., 58
Girard, G., 138
Gitomar, D., 136
Glaister, S., 414
Glover-Kudon, R., 237
Godfrey-Smith, P., 114, 215
Goldring, E. B., 56
Goodlad, J., 155
Green, B. L., 404
Greene, J., 352, 417, 424
Greene, J. C., 9, 15, 17, 180, 192, 222, 224, 225, 288, 318, 321, 385, 386, 485, 488, 495
Gregersen, H. M., 412
Greiner, J. M., 425
Guba, E. G., 47, 49, 98, 115, 116, 191, 197, 198, 301, 344, 367, 392, 420
Gunn, E. M., 362

Hall, B. L., 117, 201
Halpern, E. S., 98
Hambrick, R., 16
Hammond, R. L., 158
Hanson, G. S., 266
Hanssen, C. E., 370, 371
Hastings, J. T., 165
Hatry, H. P., 269
Hebert, Y. M., 392
Hendricks, M., 29, 51, 58, 83, 236
Henry, G., 37, 152
Henry, G. T., 6, 7, 9, 14, 15, 16, 23, 224, 350, 385, 396, 405, 408, 428, 445, 455, 473, 480, 481, 483, 484, 486, 495
Heron, J., 203
Hill, W. H., 42
Hodgkinson, H., 304
Holly, F. R., 458
Holupka, C. S., 16
Honea, G. E., 82, 94
Hood, S., 233
Hood, S. L., 98, 234, 235
Hook, B. B., 222
Hopson, R., 233
Horst, P., 268
House, E., 122, 185, 196, 215, 216, 217, 218, 219
House, E. R., 15, 17, 115, 116, 143, 264, 287, 301, 325, 395, 463, 493
Howe, K. R., 15, 17, 115, 116, 117, 185, 215, 216, 217, 218, 219, 264, 287, 325
Hoxby, C. M., 56
Hrabik, L., 279
Huang, X., 433
Huberman, A. M., 98, 317, 485
Huffman, L. C., 233
Hughes, E. F. X., 224, 480
Hughes, S., 233
Humphreys, L., 91
Hurst, J., 304
Hurworth, R., 420
Huxley, E., 198

Imm, P. S., 211
Impara, J. C., 138
Introcaso, D., 51

Johnson, E. C., 395
Johnson, T. D., 222
Joint Committee on Standards for Educational Evaluation, 7, 8, 11, 96, 97, 240, 267, 357, 367, 369, 449, 457, 479
Jordan, G. B., 293
Jorgensen, D. L., 424
Julnes, G., 6, 7, 9, 14, 15
Jun, S., 118, 324

Kaftarian, S., 209, 211
Kaftarian, S. J., 406
Kahn, J. P., 91
Kaltreider, L., 266
Kane, M., 138
Kaplan, A., 120, 246
Kaplan, G., 56
Karlsson, O., 218
Kay, P., 268

Kee, J. E., 414
Kellaghan, T., 45
Kemmis, S., 203
King, J. A., 12, 14, 199, 200, 222, 238, 277, 321, 350
Kirkhart, K. E., 98, 232, 481, 482
Kirkpatrick, D. L., 20, 334
Kirkpatrick, J. D., 334
Kline, R. B., 445
Knott, T. D., 94
Krathwohl, D. R., 42
Krueger, R. A., 438, 439
Kuhn, T. S., 215
Kushner, S., 219
Kvale, S., 434, 437

Lambert, F. C., 184
Lambur, M. T., 29, 30
LaVelle, J. M., 49, 492
Lawrenz, F., 370
Layard, R., 414
Leahy, J., 279
Lee, A. M., 458
Leeuw, F. L., 300
Leithwood, K. A., 462, 480, 485
Levin, H. M., 411, 412, 413, 414
Levine, H., 304
Leviton, L. C., 29, 45, 52, 66, 75, 76, 120, 143, 144, 196, 224, 293, 480, 485
Lewin, K., 12
Light, R. J., 335, 381
Lincoln, Y. S., 47, 49, 98, 115, 116, 117, 197, 198, 301, 344, 367, 392, 420, 437
Linton, R., 318
Lipsey, M. E., 40, 164
Lipsey, M. W., 17, 292, 293, 384, 404, 407
Logan, T. K., 266
Love, A. J., 29, 274, 275
Lovell, R. G., 102
Lyman, D. R., 266

Mabry, L., 104, 441
MacDonald, B., 215, 216
Mackay, K., 237
MacNeil, C., 222
Madaus, G., 45, 164
Madaus, G. F., 155, 165
Madison, A., 235
Mark, M. M., 6, 7, 9, 14, 15, 37, 66, 69, 160, 181, 201, 224, 350, 396, 405, 455, 480, 481, 483, 484, 492, 493, 494
Martin, C., 266
Masia, B. B., 42, 165
Massey, J. G., 412
Mastroianni, A. P., 91
Mathison, S., 11, 102
Mattox, J. R., 71
McCall, R. B., 404
McClintock, C., 301
McCombs-Thornton, K., 16
McCracken, G., 434, 437
McEwan, P. J., 411, 412, 414
McLaughlin, J. A., 293
McTaggert, R., 203
Meier, K. J., 362
Merriam-Webster, 67
Mertens, D. M., 17, 50, 116, 201, 234, 277, 350
Milanowski, A., 324
Miles, M. B., 98
Milgram, S., 91
Milich, R., 266
Miller, R., 46, 66, 196
Miller, R. L., 212, 213
Miller-Stevens, K., 137
Mills, A. S., 412
Milstein, B., 237
Minnema, J., 12, 277
Moe, T. M., 56
Montrosse, B. E., 71
Morrell, J. A., 147
Morris, M., 79, 80, 81, 88, 94, 95
Mueller, M. R., 306
Muñoz, M. A., 56
Murphy-Berman, V., 325
Murray, F. B., 135, 136

Nachmias, D., 115
Nahan, N., 290, 300
National Performance Review, 165
Nay, J. N., 268
Neal, A. D., 134
Newcomer, K. E., 269, 496
Novak, S. P., 266
Nowakowski, J., 159

O'Brien, P. M., 133, 134, 136
O'Sullivan, E., 394, 407
O'Sullivan, R. G., 199
Oakes, J. M., 91, 92
Oh, C. H., 481
Osborne, D., 57
Oswald, F. L., 433
Owen, J. M., 184

Padilla Stone, S., 222
Patton, M. Q., 6, 14, 16, 17, 25, 30, 32, 49, 59, 70, 117, 179, 180, 181, 185, 201, 207, 208, 212, 214, 222, 224, 235, 240, 316, 321, 322, 346, 370, 371, 391, 437, 446, 455, 461, 480, 481, 485
Pawson, R., 167, 445
Perrin, B., 57
Perry, J., 118, 324
Perry, K. M., 370
Peters, G. B., 8
Piontek, M., 455, 456, 459, 480
Plantz, M. C., 29, 51
Posavac, E. J., 323
Positer, T. H., 165, 184
Preskill, H., 14, 17, 25, 49, 59, 60, 184, 218, 238, 239, 264, 313, 455, 456, 459, 480
Pritchard, K. J., 29, 51
Provus, M. M., 155

Radin, B., 57
Rallis, S. F., 17, 30, 115, 117
Rassel, G. R., 394, 407
Reason, P., 203
Reichardt, C. S., 59, 115, 117, 208, 404
Resnick, L. B., 54
Riccio, J., 37, 188, 416
Riccio, J. A., 339
Rich, R. F., 481
Rist, R. C., 44, 50, 53
Rodosky, R. J., 56
Rodriguez-Campos, L., 199
Rog, D. J., 16, 32, 183, 339, 376, 417
Rogers, E. M., 481
Rogers, P. J., 59, 292
Rosenbaum, D. P., 266
Rossi, P. H., 39, 40, 42, 160, 296
Rossman, G. B., 17, 30
Rouge, J. C., 82, 83
Rubin, H. I., 435, 437

Rubin, I. S., 435, 437
Ryan, C. S., 404
Ryan, G. W., 446
Ryan, K. E., 218, 219, 222

Sackett, D. L., 384
Salkind, N. J., 445
Sanders, J., 29, 52, 53, 55, 184, 215, 236
Sanders, J. R., 46, 82, 236, 238, 496
Scanlon, J. W., 268
Schmidt, R. J., 406
Schneider, M., 56
Schnoes, C. J., 325
Schöen, 205
Schonlau, M., 433
Schwandt, T. A., 6, 7, 17, 33, 34, 98, 116, 143, 145, 218, 219, 240, 262, 384, 395, 446
Scott, A., 414
Scriven, M., 7, 8, 9, 13, 20, 21, 23, 26, 29, 33, 39, 46, 47, 102, 110, 120, 130, 144, 145, 146, 147, 167, 168, 225, 248, 288, 335, 368, 370
Sechrest, L., 65, 118, 213, 414
Sedlak, A. J., 389
Senge, P. M., 59, 205
SenGupta, S., 233, 235
Shadish, W., 46, 66, 196
Shadish, W. R., 15, 16, 45, 47, 66, 112, 120, 122, 143, 144, 196, 343, 404
Shapira, R., 56
Shih, J., 233
Shinkfield, A. J., 175, 177, 340, 341, 369
Shotland, R. L., 201
Shroter, D. C., 71
Shula, L. M., 205, 321, 332, 480, 481, 485
Sieber, J. E., 104–105
Smith, E. R., 41
Smith, I. E., 237
Smith, M. F., 182, 183, 263, 269
Smith, N. L., 90, 213, 214, 265, 318
Smith, P. V., 381
Smyth, J., 430
Society for Prevention Research, 370
Sonnichsen, R. C., 29, 32, 102, 274, 275, 276
Spiegel, A. N., 344
Springer, M. G., 118
St. Pierre, T. L., 266
Stake, R. E., 7, 46, 47, 143, 145, 156, 191, 192, 193, 194, 195, 196, 218, 234, 321, 336, 341, 368, 389, 390, 392, 434, 435, 437, 445, 446
Standaert, R., 39
Stanley, J. C., 46, 115
Stecher, B. M., 473
Stephan, A. S., 41
Stevahn, L., 12, 277
Stevenson, J., 119
Stockdill, S. H., 237
Strauss, A., 446
Stufflebeam, D. L., 7, 12, 29, 45, 46, 47, 82, 83, 111, 112, 121, 155, 159, 164, 173, 174, 175, 176, 177, 213, 218, 248, 326, 340, 341, 367, 368, 369
Suchman, E., 46, 66, 160, 480
Sullins, C., 184
Sullins, C. D., 71
Sylvia, R. D., 362

Tabachnick, B. G., 343
Talmadge, H., 13
Tarnai, J., 431
Tashakkori, A., 117, 118
Taylor-Powell, E., 160, 276, 279, 292, 293, 294
Teddlie, C., 117, 118
Teske, P., 56
Tharp, R. G., 267
Thomas, D., 119
Thompson-Robinson, M., 233
Tilley, N., 445
Torres, R. T., 14, 17, 25, 59, 184, 218, 222, 264, 455, 456, 459, 480
Toulemond, J., 66
Trochim, W. M. K., 318, 404, 493
Tyler, R. W., 41, 154

U. S. Department of Education, 136

Van Mondfrans, A., 467
Vestman, O. K., 69, 70
Vroom, P. E., 290, 300

Wadsworth, Y., 235
Wandersman, A., 209, 211, 212, 222, 237, 326, 335, 350
Weaver, L., 222, 224
Weiss, C. H., 15, 18, 44, 46, 66, 69, 70, 95, 153, 160, 162, 165, 181, 205, 224, 264, 292, 293, 296, 297, 323, 405, 454, 462, 480, 481
Wennberg, J. E., 384
Wertz, F. J., 473
Wetterhall, S. F., 237
White, B. F., 45, 46
White, K. R., 348
Whitmore, E., 199, 201, 202, 203, 209, 317
Wholey, J. S., 17, 44, 45, 46, 57, 182, 183, 184, 268, 269
Williams, A., 414
Williams, B., 59
Wingate, L. A., 12
Winston, J., 57
Winters, M. A., 118
Witkin, B. R., 263
Wolfe, E. W., 433
Worthen, B. R., 46, 160, 248, 252, 264, 277, 348, 491

Yates, B. T., 414
Yin, R. K., 343, 389, 392, 404, 406, 446

Zimmerman, R., 266
Zvoch, K., 426, 445

Subject Index

Absolute standards, defined, 334–335
Abstracts, reports, 471
Accountability
  CIPP model, 173–179
  defined, 11
  focus on, 53–59
  outcomes, measurement of, 58
  politics of evaluation, 69
  stages of theory evaluation, 122
  standards for, 84–85, 501
Accreditation
  associations for, 40, 129–137
  neutrality, transparency and purpose, 135–136
  politics of, 133–134
  process of, 132–134
  UCLA model, 179
Accuracy, 11, 84–85, 96–97, 500
Achievement tests, history of, 40. *See also* standardized testing
Action hypothesis, 296–298
Action research approach, 12–13, 203
Ad hoc individual reviews, 128, 139
Ad hoc panel reviews, 128, 137–139
Administering evaluations. *See* conducting evaluations
Administrators, 173. *See also* stakeholders
Adult education, 201, 205–207. *See also* education
Advocacy, 210, 213
African Evaluation Association, 50–51, 82–83
Agreements and contracts, 367–368
Alternative views of evaluation, overview, 111–117
Amazon.com, 147
American Bar Association, 132
American Cancer Society, 237
American Educational Research Association (AERA), 149
American Education Research Associations Division H, 48
American Evaluation Association (AEA). *See* Guiding Principles for Evaluation (AEA)
  current trends, 49
  history of, 48
  metaevaluations, 369–371
  Organizational Learning and Evaluation Capacity Building, 238
  political environment, 75
  purpose of evaluation, 15
  testing in education, response to, 55
  website, 90
American National Standards Institute (ANSI), 83
American Psychological Association (APA), 149, 369–371
Analysis, CIPP model, 175
Annie E. Casey Foundation, 160
Annual measurable objectives (AMOs), 165
Anonymity, 92, 353
Anti-poverty programs, 43–46
Appendices, reports, 475–476
Apple, Inc., 147
Approaches, defined, 112, 244
Asia, ethical codes, 82
Aspen Institute, 165, 292
Assessment, 439–442. *See also* standardized testing
Association of American Colleges and Universities, 134
Association of American Medical Colleges, 132
Audiences, for evaluation. *See also* stakeholders
  defined, 260
  identifying, 287–290
  oral reports, 476–478
  reports, planning for, 354–357
  report writing, 458–461
Audiovisual presentations, 476–478
Auditing evaluations, 276
Audit trail, 98–99
Australasian Evaluation Society, 49–51, 82
Australia, 57, 292
Authenticity, 197
Autocratic evaluator, 216

Behaviors, observation of, 423–427
Belmont Report (1979), 91
Bias
  accreditation, 135–136
  democratically oriented approaches, 215, 218
  empowerment approach, 213–214
  evaluation personnel and, 306–307
  financial relationships and, 100–102
  interpersonal relationships and, 99–100
  methodology, choice of, 117–120
  organizational relationships and, 102–104
  overview, 94–98
  participant-oriented approaches, 225
  program-oriented approaches, 167
  reducing, 98–99
  reports and, 457

survey data, 434
theory-based approaches, 161
Bickman, Leonard, 160–164
Blogs, reports and, 460–461
Blue-ribbon panels, 127–128, 138–139
Boundaries, setting
approach used, variations in, 309–310
decision to proceed, 310–311
evaluation object sample, 302–304
identifying stakeholders and audiences, 287–290
overview, 290–292
perceptions, differences in, 300–301
political contexts, 307–308
program changes, managing, 301
program description, methods for, 298–300
program theory, development of, 298–300
program theory and logic models, use of, 292–298
resource analysis, 304–307
Branching, survey questions, 431–432
Budget, evaluation management
cost-benefit analysis, 411–413
cost-effectiveness studies, 413–414
planning, 358–366
survey costs, 432
Budgetary decision making, 58
Bureaucratic evaluator, 216
Buros Institute of Mental Measurements, 149
Bush, George H. W., 54, 58
Business, use of evaluation, 19

California Achievement Test, 440
Canada
accreditation, education, 131
evaluation profession, 50
performance monitoring, 57
Canadian Evaluation Society (CES)
current trends, 49
ethical code, 82
metaevaluations, 369–371
Capacity building approaches
data collection methods, 350
empowerment approach, 209–215
limitations to, 240–241
overview, 231
role of evaluation in, 235–241
Capstone projects, 137
Case/context tradition, 119
Case studies
data collection, 343–344, 388, 389–392, 404–406
participant-oriented approaches, 197–199
purposive sampling, 409–411
responsive approach, 194
Causal designs, data collection, 343–344, 388, 394–400, 404–406
Causal hypothesis, 296–298
Causal relationships
constructivist paradigm, 118
experimental tradition, 119
postpositivism, 115
program descriptions, 296–298
stages of evaluation theory, 120
Centers for Disease Control and Prevention, 237
Central tendency statistics, 354
Certification, UCLA model, 179
Chain of evidence, 405
Charter schools, 56
Charts, reports, 473
Checklists
contracts, 367
decision to proceed, 311
educational products, 146, 147
evaluation criteria, sources for, 324
external evaluator, use of, 276
Key Evaluation Checklist, 370
observations, data collection, 426–427
potential stakeholders and audiences, 289
reports, 479
sampling procedures, 410
selecting evaluators, 280–281
when to conduct an evaluation, 271, 272
Chen, Huey, 160–164
CIPP model. *See* Context-Input-Process-Product (CIPP)
Clarity, standards for, 86
Clash of cultures, 71–75, 72, 81

Clients. *See also* stakeholders
communication with, 75–78
data collection, planning for, 347, 350
defined, 260
negative messages, delivering, 466–469
reports, 485–486
Clinton, William J., 54, 57
Clock model, responsive approach, 194–195
Collaborative evaluation, 199
Colorado Department of Education, 136
Commission on the Future of Higher Education, 133, 138–139
Communication
about causal designs, 394–395
CIPP model, 175–176
cost of, 365
data collection design selection, 345
establishing and maintaining, 75–78
ethical concerns, 80
evaluation management, 358–366
importance of, 65
planning for, 261, 354–357
political environment, 74, 75
public's right to know, 216
reports
accuracy, balance and fairness, 457
appearance, 465–466
human considerations, 466
key components, 469–476
methods for, 455–456
negative messages, 466–469
oral reports, 476–478
planning for, 354–357
purposes of, 454–455
review of, 468–469
strategies for, 463–464
tailoring to audiences, 458–461
technical writing style, 464
timing of, 461–463
uses of information, 479–486
standards about, 96–97
technology resources, 307

Comparative analysis of approaches
  choice of approach, 247
  contributions of approaches, 248
  mixed use of approaches, 244–245
  model synthesis, calls for, 245–247
  summary of, 249–251
  uses of approaches, 249–253
Comparison group designs, 388, 403–404
Competence, standards for, 87, 88, 501
Competition
  in education, 56
  politics of evaluation, 67–71
Comprehensive Test of Basic Skills, 440
Computer costs, 365
Conceptual use of findings, 25, 480
Conclusions
  cultural competence, 232
  reports, presentation of, 474–475
Conducting evaluations
  data analysis and interpretation, 444–449
  data collection
    case studies, 389–392
    causal designs, 394–400
    cost analysis, 411–414
    cross-sectional designs, 392–393
    descriptive designs, 387–394
    existing documents and records, 419–420
    experimental designs, 397–400
    focus groups, 437–439
    interviews, 434–437
    mixed methods, using, 383–386, 405–407
    observations, 423–427
    overview, 381–383
    planning and organizing, 343–345, 443–444
    quasi-experimental designs, 400–407
    sampling methods, 407–411
    sources and methods, identifying, 345–350, 420–423, 442
    surveys, 427–434
    tests and assessments, 439–442
    time-series designs, 393–394
  planning for
    agreements and contracts, 367–368
    data collection, 343–350
    management plan, 358–366
    metaevaluations, 368–374
    organizing and evaluating data, 353–354
    overview, 340–343
    reports, 354–357
    sampling and procedures, 351–353
    summary worksheet, 355–357
Confidentiality, 80, 86, 92–93, 136, 353
Confirmability, 197
Conflict of interest
  accreditation, 135–136
  democratically oriented approaches, 218
  empowerment approach, 213–214
  evaluation personnel and, 306–307
  financial crisis 2008, 78–79
  financial relationships and, 100–102
  interpersonal relationships and, 99–100
  organizational relationships and, 102–104
  overview, 94–98
  participant-oriented approaches, 225
  program-oriented approaches, 167
  reports, 457
  standards for, 86
  theory-based approaches, 161
  types of problems, 79–82
Connoisseurship and criticism, 139–141, 144–145
Consent, informed, 86, 91–94, 347
Constructivism paradigm
  background and preference, 118
  overview, 115–116
  participant-oriented approaches, 191, 197–199
  practical participatory evaluation, 205–207
Consultants, evaluation costs, 364–365
Consumer-oriented evaluations
  application of, 145–147
  comparative analysis, 249
  developers of, 144–145
  heuristics of, 322
  influences of, 147–149
  overview, 123, 143–149
  planning evaluations, 281–283, 309
Consumers' Union, 144
Content, reports, 458–460
Content analysis procedures, 420
Content standards, 165
Context. *See also* political considerations
  approach used, variations in, 309–310
  decision to proceed, 310–311
  evaluation object sample, 302–304
  perceptions, differences in, 300–301
  planning evaluations, 262
  political contexts, planning for, 307–308
  program changes, managing, 301
  of reports, planning for, 354–357
  resource analysis, 304–307
  stakeholders and audiences, identifying, 287–290
  tradition of, 119
Context-Input-Process-Product (CIPP)
  history of, 47
  overview, 123, 173–179
  strengths and limitations, 184–185
  uses of, 184
Continuous improvement, 25, 154–155, 211
Contracts, 79–80, 85–86, 87, 367–368
Control of decisions, 206, 210
Control of evaluation, 200–204

Convergent phase, question selection, 315, 328–332
Cooperative inquiries, 203
Cost, evaluation management
  cost-benefit analysis, 411–413
  cost-effectiveness studies, 413–414
  planning, 358–366
  surveys, 432
  use of existing documents and records, 419–420
Cost-benefit analysis, 157, 263, 411–413
Cost-effectiveness studies, 263, 413–414
Cost-free evaluation, 304
Countenance Framework, 191–192, 195–197
Cousins, Bradley, 200, 205–207
Credibility
  cultural competence, 232
  participant-oriented approaches, 197–199, 224–225
  political environment, 73
Criteria. *See* evaluation criteria, identifying and selecting
Criterion-referenced testing, 165, 439–442
Criticism, as evaluation, 139–141
Criticism, of evaluation, 77–78, 468–469
Cronbach, Lee, 121
Cross-sectional designs, data collection, 344, 388, 392–393, 406, 407–408
Cultural competence
  bias and, 98
  data collection and, 93–94
  empowerment approach, 211
  role of, 232–235
  standards for, 88, 89
Cultural considerations. *See also* culture, organizational
  bias, 97–98
  global evaluation profession, 50–51
  political environment, working within, 71–75
Culture, organizational
  causal designs, use of, 396
  evaluation impact on, 59–61
  practical participatory evaluation, 205–206
  report findings, use of, 481

Data analysis and interpretation, 444–449. *See also* reports
Databases, as information source, 346–347
Data collection
  case studies, 389–392, 404–406
  causal designs, 394–400
  CIPP model, 175, 178
  comparison groups designs, 403–404
  cost analysis, 411–414
  cross-sectional designs, 392–393
  cultural competence and sensitivity, 93–94
  descriptive designs, 387–394
  ethical concerns, 81, 89
  existing documents and records, 419–420
  experimental designs, 397–400
  feasibility and propriety, 398–400
  focus groups, 437–439
  interrupted time-series design, 400–403
  interviews, 434–437
  logical positivism, 114–115
  mixed methods, using, 383–386, 405–407
  observations, 423–427
  organizing and evaluating, 353–354
  overview, 381–383
  ownership of data, 216
  planning for
    collection methods, 348–350, 407–411, 442
    design selection, 343–345
    identifying sources, 345–347, 420–423
    potential problems, 443–444
    sampling and procedures, 351–353, 407–411
    summary worksheet, 355–357
  political environment and, 74–75
  posttest-only design, 397–398
  pre-post designs, 398
  quasi-experimental designs, 400–407
  regression-discontinuity design, 404
  standards for, 85
  surveys, 427–434
  tests and assessments, 439–442
  time-series designs, 393–394
Decision-based approaches, evaluation. *See also* Utilization-Focused Evaluation (UFE)
  CIPP model, 173–179
  comparative analysis, 250
  developers, 173–177
  evaluability assessment, 182–184
  heuristics of, 322
  overview, 123, 172–173
  performance monitoring, 182–184
  planning evaluations, 281–283
  strengths and limitations, 184–185
  UCLA model, 178–179
  uses of, 184
Decisions, control of, 206
Decisions, cultural competence and, 232
Democracy, evaluation purposes, 25
Democratically oriented approaches
  identifying stakeholders and audiences, 287–288
  overview, 123, 204, 215–219
  use of results, 264
Democratic evaluator, 216
Denmark, 50
Dependability, 197
Description, Contenance Framework, 191–192, 195–197
Descriptive documents, 298
Descriptive evaluations, 344–345, 387–394
Design phase, evaluation. *See* planning evaluations
Development, international, 116–117, 199–200, 201, 203

Developmental approaches, 14, 123, 202–204, 207–208
Dialogue, use of, 218
Dillman, Donald, 430
Direct uses of evaluation, 263
Disclosure, standards for, 86
Disclosure agreements, 80
Discrepancy Evaluation Model (DEM), 155–158
Disenfranchised groups, 217
Dispersion descriptive statistics, 354
Displays, reports, 473
Divergent phase, question selection
  defined, 315
  different approaches as heuristics, 321–323
  evaluator judgments, 325–326
  expert consultants, as source, 325
  multiple sources, summarizing, 326–328
  overview, 315–316
  research and field work as source, 323–324
  stakeholder needs, identifying, 316–321
Diversity, 88, 232–233. *See also* cultural competence
Doctoral examinations, 127
Documentation. *See also* communication; data collection; reports
  bias, reducing, 98–99
  descriptive documents, planning, 298
Domain-referenced testing, 439–442
Donaldson, Stewart, 160–164
Driving under the influence example, 302–304

Eastern Europe, 50, 67
ECB Theory of Change, 293–294
Economic considerations, transformative paradigm, 116–117
Education
  accountability, focus on, 53–59
  accreditation process, 129–134
  adult education, 201
  case/context tradition, 119
  CIPP model, 184
  conceptions of evaluation, 113–114
  connoisseurship and criticism, 139–141
  Elementary and Secondary Education Act (1965), 44–45
  of evaluation personnel, 305–307
  history of, 1800-1940, 39–41
  history of, 1940-1964, 42–43
  internal evaluations, 52
  objectives-oriented approaches, 154–159
  program-oriented approaches, 165
  results, used for educating others, 263
  use of evaluation in education, 18–19
  What Works Clearinghouse, 148
Efficiency, 41
Eight Year Study, 41
Eisner, Elliot, 128–129, 139–143
Electronic interviews and surveys, 442
Electronic reports, 460–461
Elementary and Secondary Education Act (1965), 44–45, 54, 154–155
E-mail surveys, 432–433
Empowerment approaches
  democratically oriented approaches, 215
  evaluation personnel training, 306–307
  overview, 123, 202–204, 209–215
  purpose of evaluation, 25
  use of results, 264
Enabling objective, defined, 157
Encyclopedia of Evaluation, 199
England, 82
Enron Corporation, 79
Epistemology, 114–117
Equality, democratically oriented approaches, 217
Equipment, cost of, 365–366
Equity, question selection, 317–318

ESEA. *See* Elementary and Secondary Education Act (1965)
Ethics
  agreements and contracts, 367–368
  beyond a code of ethics, 104–105
  bias and conflicts of interest, 94–100, 102–104
  data collection, 399, 443
  financial relationships, bias and, 100–102
  guidelines, development of, 48
  human subjects, protections for, 90–94
  overview, 64–65
  program-oriented approaches, 166
  propriety of evaluation, 267–268
  standards, overview, 78–79, 82–83
  types of problems, 79–82
Ethnicity, cultural competence, 88
European Commission, 67
European Evaluation Society, 49–51
European Union, 50, 67
Evaluability assessment, 182–184, 268–271
Evaluability assessments, 123
Evaluation
  alternative views, overview, 111–112
  classification of, 120–124
  comparative analysis
    choice of approach, 247
    contributions of approaches, 248
    mixed use of approaches, 244–245
    model synthesis, calls for, 245–247
    summary of, 249–251
    uses of approaches, 249–253
  conceptions of, 113–114
  developmental, 202–204
  disciplinary boundaries and methodology, 119–120
  formative and summative, overview of, 20–25
  future of, 490–496

graduate programs, 46–47
importance and limitations, 32–35
inappropriate uses of, 265–268
informal *vs.* formal, 5–6
key terms, 6–9
mainstreaming, 236–237
methodological background and preference, 117–120
organizational learning and, 59–61
participatory, 202–204
philosophical and ideological differences, 114–117
as a profession, 47–49
purposes of, 13–16, 25
roles of evaluators, 16–18
uses of, overview, 18–20, 263–265
vision for, 496–497
vs. research, 9–13
Evaluation capacity building (ECB), 237–241, 322, 480–481
Evaluation Capacity Development Group (ECDP), 238
Evaluation Center, Western Michigan University, 158–159, 177
Evaluation criteria, identifying and selecting
convergent phase, 328–332
divergent phase
different approaches as heuristics, 321–323
evaluator judgments, 325–326
expert consultants, as source, 325
multiple sources, summarizing, 326–328
overview, 315–316
research and field work as source, 323–324
stakeholder needs, identifying, 316–321
flexibility, need for, 336
overview, 314–315
specifying criteria and standards, 332–335
standards, checklists and guidelines, 324
Evaluation cube, 158–159
Evaluation Inquiry for Learning in Organizations (1998), 59
Evaluation request and responsibilities
differences in approaches, 281–283
evaluability assessment, 268–271
hiring evaluators, 277–281
internal *vs.* external evaluator, 271–276
overview, 259–260
purpose of evaluation, 260–265
Evaluation Research Society and Evaluation Network, 48
Evaluators, roles of, 16–18, 210
Executive summaries, 470–471
Experimental designs, 397–400, 406
Experimental tradition, 119
Expert, evaluator as, 17
Expert consultants, 325
Expertise-oriented evaluations
accreditation, 129–137
ad hoc individual reviews, 139
ad hoc panel reviews, 137–139
comparative analysis, 249
connoisseurship and criticism, 139–141
developers of, 128–129
heuristics of, 322
influence of, 142–143
informal review systems, 137
overview, 123, 127–128
planning evaluations, 281–283, 309–310
External evaluation, overview, 27–32
External evaluators
bias and, 102–104
decisions about, 271–276
needs and assignments, 363–364
External review, metaevaluations, 371

Face-to-face surveys, 433–434
Facilitation, empowerment evaluation, 210
Facility costs, 366
Fact-value dichotomy, 116
Fairness, standards for, 86
Feasibility
data collection, 398–400
defined, 11
participant-oriented approaches, 224–225
standards for, 84–85, 500
Federal government. *See* government programs
Feedback, of evaluation, 77–78
Fetterman, David, 157–158, 209–215
Financial aid, education, 133–134
Financial crisis of 2008, 78–79, 95
Financial disclosure, standards for, 96
Financial relationships, bias and, 100–102
Financial resources, 304–305
Fiscal responsibility, standards for, 86
Fitch rating agency, 78–79, 95
Flexibility, responsive evaluation, 193
Focus, of evaluation, 174–175, 472
Focus groups, 386, 437–439, 442
Follow-up studies, 157
Formal agreements, 85–86, 87, 232
Formal evaluation, 5–6
Formal professional review systems, 128, 129–137
Format, reports, 458–460, 465–466
Formative evaluations
accreditation, 134
appropriate use of evaluation, 267
bias and, 104
CIPP model, 175
evaluability assessments, 183
foundations and nonprofit organizations, 48–49
internal *vs.* external evaluator, 271
metaevaluation, 368–374
overview, 20–25
purpose of, 304
reports, 454–455, 461, 462, 474–475
Foundations, 48–49, 58–59
Fourth-Generation Evaluation (1989), 197–199
Fourth National Incidence Study of Child Abuse and Neglect, 389
France, 50, 82
Functional analysis, 145–147

Funding agency review panels, 138
Funding decisions, theory-based approaches, 161
Future trends, evaluation, 490–496

Gantt charts, 361–362
Gender, cultural competence, 88
General Accounting Office (GAO), 43–44
General welfare, standards for, 15, 87, 89–90, 503–504
Germany, 50, 82, 131
Getting to Outcomes (GTO), 211
Goal-free evaluation, 47, 167–169
Goals, 13–16, 42, 154–159, 211, 225. *See also* objectives-based approaches; program-oriented approaches
Goals 2000: Educate America Act, 54
Gore, Al, 57
Government Accountability Office (GAO), 43–44, 69
Government Performance Results Act (GRPA), 44, 57, 165, 292
Government programs
  during 1940-1964, 42–43
  during 1964-1972, 43–47
  accreditation, educational institutions, 133–134
  decision-oriented approaches, 184
  Great Society programs, 43–46
  New Deal programs, 41
  performance monitoring, 56–58, 182–184
  politics of evaluation, 67–71
  program-oriented approaches, 165
  use of evaluation, 19
Graduate programs, evaluation, 46–47, 492
Grant requirements, 264
Graphics, reports, 465–466, 473, 478
Great Depression, 41
Great Society, 43–46
Guardian permission, 347
Guba, Egon, 176, 197–199
Guidelines, evaluation criteria, 324
Guiding Principles for Evaluation (AEA)
  agreements and contracts, 367–368
  cultural competence, 232
  data interpretation, 449
  history of, 48
  metaevaluations, 369–371
  political environment, 75
  purpose of evaluation, 15
  reports, 457
  stakeholder expectations and understandings, 394–395
  text of, 501–504
  value of, 244
Guiding Principles for Evaluators (AEA), 82, 87–90, 96–97

Hamilton, Alexander, 128
Head Start, 46
Heuristics, evaluation approaches as, 321–323
Higher education, accreditation, 130–132
Higher Education act, 133–134
High stakes testing, 55. *See also* standardized testing
Hiring evaluators, 277–281, 363–364
History of evaluation
  1800-1940, 39–41
  1940-1964, 42–43
  1964-1972, 43–47
  1990-present, 49–61
  accreditation process, 129–130
  early forms, 38–39
  summary of, 61
Honesty, standards for, 87, 88, 96, 502
Howe, Kenneth, 216
Human rights and respect standards for, 89
Human rights and respect, standards for, 86
Human subjects, protections for, 90–94, 347, 443
Huxley, Elspeth, 198
Hypotheses, program descriptions, 296–298

IASA (Improving America's Schools Act), 54
Ideological differences, evaluation, 114–117
Immediate program impacts, logic models, 160
Impact, of evaluation, 15
Impact studies
  data collection, 344
  empowerment approach, 211
  evaluation questions, 388
  focus on, 53–59
  planning evaluations, 263
Impartial Reporting standard, 96–97
Implementation, manageability of, 222
Implementation failure, 161
Improvement/accountability approach, 121–124
Improving America's Schools Act (1994), 54
India, 131
Indirect uses of evaluation, 263–265
Industry, use of evaluation, 19
Inequalities, responsive evaluation, 196. *See also* empowerment approaches
Influence, report findings, 481–486
Informal evaluation, 5–6
Informal professional review systems, 128, 137
Informants, role of, 198
Information, collecting. *See* data collection
Informational uses of evaluation, 263
Informed consent, 86, 91–94, 347, 353
Inputs, CIPP model, 173–175
Inputs, logic models, 159–160
Installation, Discrepancy Evaluation Model, 156
Institutionalization, empowerment approach, 211
Institutional Review Boards (IRBs), 90–94, 347
Instrumental use of findings, 480
Instrument fallacy, law of, 119–120
Integrity, standards for, 87, 88, 96, 502
Intention, 482
Interest groups. *See* stakeholders
Interest profiles, history of, 40

Interim reports, 455, 462. *See also* reports
Internal data sources, 346–347
Internal evaluations, overview, 27–32
Internal evaluators
  bias and, 102–104
  decisions about, 271–276
  interpersonal relationships and, 99–100
  needs and assignments, 363–364
  skills of, 51–53
Internal review, metaevaluations, 371
International development work, 116–117, 199–200, 201
International evaluation, 49–51
International evaluation, future of, 492
International Institute on Administrative Sciences (IIAS), 44
International Organization for Cooperation in Evaluation (IOCE), 50–51
Internet
  reports, creating, 460–461
  surveys on, 430, 432–433
Interpersonal considerations, 64–65, 75–78, 99–100. *See also* ethics; political considerations
Interrupted time-series designs, 388, 400–403
Intersubjectivity, postpositivism, 115
Intervention hypothesis, 296–298
Interviews. *See also* surveys
  data collection process, 434–437
  face-to-face surveys, 433–434
  hiring evaluators, 280
  mixed methods research, 386
  overview, 442
  participant-oriented approaches, 197–199
  program planning, 298–299
  purposive sampling, 409–411
  sampling methods, 352
  with stakeholders, 316–321
  telephone surveys and interviews, 431–432
Introduction, reports, 471–472
Iowa Test of Basic Skills, 440

Japan, 50n
Johnson, Lyndon, 43–44
Joint Committee on Standards for Educational Evaluation
  agreements and contracts, 367–368
  bias and conflict of interest, 96–97
  cultural competence, 232–233
  data interpretation, 449
  history of, 48
  metaevaluation, 369–371
  political environment, 71
  Program Evaluation Standards, 75, 82–87, 244, 499–501
  propriety of evaluation, 267–268
  stakeholder expectations and understandings, 394–395
Journals, evaluation profession, 48, 493
Journals, peer review process, 127, 137
Judgment, 68, 123, 191–192, 195–197

K-12 education, 132–134. *See also* education
Kennedy, Robert F., 45
Key Evaluation Checklist (KEC), 147
Kirkhart, Karen, 232
Knowledge, purpose of evaluation, 15–16
Knowledge construction, 205–207

Law of instrument fallacy, 119–120
Legal rights, standards for, 86
Lincoln, Yvonna, 197–199
Local government. *See* government programs
Local knowledge, responsive evaluation, 193
Logical positivism, 114–115
Logic models
  overview, 153–154, 159–160
  planning evaluations, 261, 309
  program description, creation of, 292–298
  strengths and limitations, 166–169
Long-term outcomes, Discrepancy Evaluation Model, 157
Long-term outcomes, logic models, 159–160

Madison, James, 128
Magnet schools, 56
Mainstreaming evaluation, 236–237, 240–241
Management by Objectives (MBO), 165
Management plan, of evaluation study, 358–366
Managers, as evaluators, 51–53
Mann, Horace, 39
Market theory, in education, 56
McNamara, Robert, 165
Medical schools, accrediting, 40, 129–130
Mental Measurements Yearbook, 149, 428, 442
Merit, 47, 173–179
Metaevaluation, 99, 368–374
Methodology. *See also* specific approach names
  background and preference, 117–120
  stages of theory evaluation, 122
Middle States Association of Colleges and Schools, 131n
Military bases, decisions about, 142
Minority reports, 475
Misinterpretation of results, 80
Mission-based approach, 132, 211
Misuse of results, 80
Mixed methods research, 385, 405–407, 422–423
Models, defined, 112, 244
Moody's rating agency, 78–79, 95
Morrell, Jonathan, 147
Multi-Attribute Consensus Reaching (MACR) procedures, 277
Multicultural validity, 232. *See also* cultural competence

National Assessment of Educational Progress (NAEP), 154
National Council for Accreditation of Teacher Education (NCATE), 132, 136
National Council on Measurement in Education (NCME), 149
National Defense Education Act (1958), 42
National Performance Review, 57
National Science Foundation, 92, 233
Nation at Risk, 53–54
Naturalistic approaches, 197–199
NDEA. *See* National Defense Education Act (1958)
Needs assessments
  consumer-oriented approach, 145–147
  empowerment approach, 211
  evaluation questions, 387
  overview, 25–27
  planning evaluations, 263
Negative messages, delivering, 466–469
Negotiation, political environment, 73
Netherlands, 50, 131
New Deal programs, 41
New England Association of Schools and Colleges, 131
New Public Management, 57
No Child Left Behind (NCLB) Act (2001), 54–56, 165, 334, 440
Nonequivalent control group design, 403–404
Non-evaluators, responsibilities of, 51–53
Nonprofit organizations
  formative evaluations, 48–49
  internal evaluations, 51–53
  logic model approaches, 160
  outcomes measurement, 58–59
  use of evaluation, 19
Norm-referenced tests, 40, 439–442
North Central Association of Colleges and Schools, 131n
Northwest Association of Accredited Schools, 131n
Nuremburg code, 93

Obama, Barack, 54n
Objectives. *See also* program-oriented approaches
  of evaluation, 13–16
  goal-free evaluation, 168–169
  participant-oriented approaches, 225
  Tylerian approach, 42
Objectives-based approaches
  heuristics of, 322
  overview, 123, 153–159
  planning considerations, 309
  uses of, 164–165
Objectives-referenced testing, 439–442
Objectivity, evaluation
  accreditation, 135–136
  logical positivism, 115
  participant-oriented approaches, 197–199
  politics of evaluation, 69
Observations
  overview, 423–427, 442
  participant-oriented approaches, 197–199
  program planning, 299
Office of Human Research Protections (OHRP), 90–91
Office of Management and Budget (OMB), 58–59
Ontology, 114–117
Opportunity costs, 412
Oral reports, 476–478
Organizational culture
  causal designs, use of, 396
  evaluation impact on, 59–61
  practical participatory evaluation, 205–206
  report findings, use of, 481
Organizational learning
  capacity building, 235–241
  developmental evaluation, 207–208
  evaluation impact, 25, 59–61
  practical participatory evaluation, 205–207
  role of culture, 232–235
Organizational relationships, bias and, 102–104
Organizational skills, of evaluators, 358–366
Organization of information, 175, 353–354. *See also* data collection; reports
Orientation, evaluation personnel, 306–307
Outcomes
  accreditation, 136
  Countenance Framework, 191–192
  emphasis on, 134–135
  empowerment approach, 211
  evaluation questions, 388
  evaluations, overview, 25–27
  focus on, 53–59
  goal-free evaluation, 168–169
  heuristics of approaches, 323
  logic models, 159–160
  measuring, nonprofit organizations, 58–59
  planning evaluations, 263
  program-oriented approaches, 165
Outcomes Measurement System, 58
Overhead costs, 366

Parental permission, 347
Participant-oriented approaches
  comparative analysis, 251
  contemporary approaches, 199–204
  democratically oriented approaches, 215–219
  developmental evaluation, 207–208
  empowerment evaluation, 209–215
  heuristics of, 322
  naturalistic and fourth generation approaches, 197–199
  overview, 189–191
  planning evaluations, 281–283, 310
  practical participatory evaluation (P-PE), 205–207
  responsive approach, 191–197
  strengths and limitations of, 223–226
  uses of, 220–222, 264

Participants. *See* stakeholders
Participation, depth of, 77, 201–204
PART (Program Assessment Rating Tool), 44
Patton, Michael, 173, 179–182, 207–208
Peer reviewed journals, 127, 137
People skills, 75–78, 225
Per diem costs, 364–365
Performance monitoring
  future trends, 496
  government programs, 56–58
  internal *vs.* external evaluator, 271
  overview, 123, 182–184
  planning evaluations, 263, 309
  school choice and, 56
  standards, specifying, 334
  strengths and limitations, 184–185
  uses of, 184
Performance tests, 441
Personal factor, defined, 179
Personality profiles, history of, 40
Personnel, availability and skills, 305–307, 363–364
Personnel, evaluation costs, 364
Personnel information, data collection, 347
Persuasion, reports, 463–464
PERT charts, 361–362
Phi Beta Kappa, 134
Philanthropic organizations, 58–59
Philosophical differences, evaluation, 114–117
Pilot testing, data collection, 74–75
Plan and procedures, reporting, 472–473
Planning, Programming, and Budgeting System (PPBS), 43–44, 165
Planning evaluations
  activities and functions, 341
  boundaries and contexts
    approach used, variations in, 309–310
    decision to proceed, 310–311
    evaluation object sample, 302–304
  identifying stakeholders and audiences, 287–290
  perceptions, differences in, 300–301
  political contexts, 307–308
  program changes, managing, 301
  program description, methods for, 298–300
  program theory, development of, 298–300
  program theory and logic models, use of, 292–298
  resource analysis, 304–307
  setting boundaries, overview, 290–292
communication, 76
conducting evaluations
  agreements and contracts, 367–368
  data collection, 343–353, 443–444
  management plan, 358–366
  metaevaluations, 368–374
  organizing and evaluating data, 353–354
  overview, 340–343
  reports, planning for, 354–357
  summary worksheet, 355–357
Discrepancy Evaluation Model, 155–158
evaluation request and responsibilities
  differences in approaches, 281–283
  evaluability assessment, 268–271
  hiring evaluators, 277–281
  inappropriate use of evaluation, 265–268
  internal *vs.* external evaluator, 271–276
  overview, 259–260
  purpose of evaluation, 260–265
overview, 257–258
political context, 73, 74
questions and criteria
  convergent phase, 328–332
  different approaches as heuristics, 321–323
  divergent phase, overview, 315–316
  evaluation criteria and standards, 332–335
  evaluator professional judgment, 325–326
  expert consultants, as source, 325
  flexibility, need for, 336
  multiple sources, summarizing, 326–328
  overview, 314–315
  research and field work as source, 323–324
  stakeholder needs, identifying, 316–321
  standards, checklists and guidelines, 324
Pluralism, responsive evaluation, 193
Policies, data collection and, 347
Policies, defined, 7–9
Policy influence tradition, 119
Political considerations. *See also* cultural competence; ethics
  accreditation, 133–134
  CIPP model, 176–177
  empowerment evaluation, 209–215
  identifying stakeholders and audiences, 287–288
  overview, 64–71
  planning evaluations, 262, 307–308
  report findings, use of, 481
  theory-based approaches, 161
  transformative paradigm, 116–117, 201
  trends in evaluation, 494–495
  utilization-focused evaluation, 179–182
  working environment, suggestions for, 71–75
Political polls, 392–393
Portfolios, 441
Positivism, logical, 114–115
Postponement of decisions, 264
Postpositivism, 115
Post secondary education, 132–134

Posttest-only designs, 388, 397–398
Poverty, 45–46, 201–202, 203
Power
  among stakeholders, 222
  democratically oriented approaches, 217
  politics of evaluation, 67–71
  question selection, 318
  transformative evaluations, 116–117, 201–202
  use of evaluation results, 264
Practical participatory evaluation (P-PE)
  heuristics of, 322
  overview, 123, 204, 205–207
  vs. transformative participatory evaluation, 201–202
  vs. utilization-focused evaluation, 181
Pragmatic approach, evaluation, 117
Preordinate evaluation, 195
Presentations, oral reports, 476–478
Pretest-posttest measurement, 155, 388, 398
Printing costs, 365
Priorities, evaluating, 387
Privacy rights, standards for, 86
Private schools, 56
Probability sampling, 408–409
Procedures
  reports, inclusion in, 472–473
  standards for, 85
Process, Discrepancy Evaluation Model, 157
Process evaluation
  CIPP model, 173–175
  overview, 25–27
  planning for, 263
Process use, defined, 59, 480
Product, Discrepancy Evaluation Model, 157
Product-oriented evaluations, 144–149, 173–175
Products, defined, 7–9
Professional associations, 48, 49, 127, 492–493
Professional journals, 493
Program, evaluation planning questions, 260–265, 269
Program Assessment Rating Tool (PART), 44, 58, 165
Program descriptions
  methods for, 298–300
  program theory and logic models, using, 292–298
  questions for, 387–388
  setting boundaries, overview, 290–292
Program Evaluation and Review Technique (PERT), 361–362
Program Evaluation Standards. See Joint Committee on Standards for Educational Evaluation
Program improvement, outcomes measurements, 58–59
Program inputs, logic models, 159–160
Program intents, 191–192
Program-oriented approaches
  comparative analysis, 250
  heuristics of, 321–322
  logic models, 159–160
  objectives-oriented approaches, 154–159
  overview, 123, 153–154
  planning evaluations, 281–283
  strengths and limitations, 166–169
  theory-based/driven, 160–164
  uses of, 164–165
Program planning and implementation, 17, 179, 264
Program professionals, as evaluators, 51–53
Programs, defined, 7–9
Promotion decisions, 127
Proposal review panels, 127
Propriety
  data collection, 398–400
  defined, 11
  of evaluation, 267–268
  standards for, 84–85, 96, 500
Provus, Malcolm, 155–158
Publications, professional, 48
Public documents, as information sources, 346–347
Public groups, input from, 73
Public relations, 264
Public sector, use of evaluation, 19
Public's right to know, 216
Public welfare, standards for, 15, 87, 89–90, 503–504
Purpose, of evaluation, 260–265
Purposive sampling, 352, 409–411

Qualitative research
  case studies, 391–392
  connoisseurship and criticism, 140, 143
  data analysis and interpretation, 445–446
  guidelines for, 92
  interviews, 434–437
  methodology, background and preference, 118–119
  observational data, 423–427
  participant-oriented approaches, 197–199, 200
  responsive approach, 194
  surveys, 429
  vs. quantitative, 161, 383–384, 419
Quality, conceptions of, 145, 211, 446
Quality, observational data, 426–427
Quality control, evaluation personnel training, 306–307
Quality management, history of, 57
Quantitative methods
  case studies, 391–392
  consumer-oriented methods, 144
  cross-sectional designs, 392–393
  data analysis and interpretation, 445
  methodology, background and preference, 117–119
  observational data, 423–427
  surveys, 427
  theory-based approaches, 164
  time-series designs, 393–394
  vs. qualitative, 161, 383–384, 419
Quasi-experimental designs, 343, 388, 400–407
Question and/or methods approach, 121–124
Questionnaires. *See* surveys

Questions, identifying and selecting
  convergent phase, 328–332
  divergent phase
    different approaches as heuristics, 321–323
    evaluator professional judgment, 325–326
    expert consultants, as source, 325
    multiple sources, summarizing, 326–328
    overview, 315–316
    research and field work as source, 323–324
    stakeholder needs, identifying, 316–321
    standards, checklists and guidelines, 324
  evaluation criteria and standards, 332–335
  flexibility, need for, 336
  interviews, 435–436
  overview, 314–315

Race, cultural competence, 88
Rand Corporation, 165
Randomized control trials (RCTs)
  controversy over, 383–384
  data collection, 397–400
  data collection, design selection, 343
  outcomes measurement, 58–59
Random samples, 351, 408–409
Rating agencies, 78–79, 95
Reality, views of, 215
Recommendations, making of, 68, 474–475
Records. *See* data collection
Recruiting, evaluation personnel, 363–364
Regression-discontinuity design, 404
Reinventing Government, 57
Relative standards, 335
Reliability, 197–199, 346–347
Religion, cultural competence, 88
Replication, postpositivism, 115
Reports
  accuracy, balance and fairness, 457
  appearance, 465–466
  case studies, 392
  CIPP model, 175–176
  ethical concerns, 80
  human considerations, 466
  impartial reporting standard, 96–97
  key components, 469–476
  methods for, 455–456
  negative messages, 466–469
  oral reports, 476–478
  planning for, 261, 354–357, 356
  political environment, 75
  purposes of, 454–455
  review of, 468–469
  strategies for, 463–464
  tailoring to audiences, 458–461
  technical writing style, 464
  timing of, 461–463
  uses of information, 479–486
Representative sample, 409–411
Requests for changes, ethical concerns, 81, 91
Research. *See also* specific evaluation approach names
  program planning methods, 299
  vs. evaluation, 9–13
Resource assessment, 211, 261, 304–307
Resources, competition for, 67–71, 161
Respect, standards for, 87, 89, 503
Responsibilities, planning for
  differences in approaches, 281–283
  hiring evaluators, 277–281
  internal *vs.* external evaluator, 271–276
  overview, 259–260
  purpose of evaluation, 260–265
Responsive and inclusive orientation standard, 85
Responsive approach, 47, 123, 191–197
Results. *See also* reports
  ethical concerns, 80
  organizing and evaluating data, 353–354
  political environment, 75
Résumés, hiring evaluators, 280
Review, of evaluation, 77–78
Rhetoric *vs.* reality, 268
Right to know, public, 216
Rist, Ray, 44
Rossi, Peter, 121, 160–164
Rubrics, 137

Sampling frame, defined, 408
Sampling methods
  overview, 351–353, 407
  purposive sampling, 409–411
  random sample, 408–409
  sample size, 407–408
School choice, 56
Scientific expert, evaluator as, 17
Scientific management movement, 41
Scope of study, standards for, 85
Scriven, Michael, 120, 144–149
Secondary education, 44–45, 132–134
Self-determination, empowerment evaluation, 209–215
Self-monitoring, 210
Self-reflection, by evaluator, 98–99
Self-study reports, 132
Sesame Street, 46
Site visits, 127, 423–427
Skill development
  of evaluation personnel, 363–364
  as use of evaluation, 263–264
Skills testing, 441
Social agenda/advocacy approach, 121–124
Social change. *See* transformative approaches
Social considerations, 116–117
Social justice
  democratically oriented approaches, 215–219, 217
  responsive evaluation, 196
  transformative evaluations, 203
Social science research
  non-evaluators, use of, 51–53
  theory-based approaches, 161
Social services, 41, 45–46. *See also* government programs
Society for Prevention Research, 324, 370
Socioeconomic status, cultural competence, 88
Software, data analysis, 446
Source of influence, 481–482

South America, 82
Southern Association of Colleges and Schools, 131n
Spellings, Margaret, 133–134
Spellings Commission, 133–136
Sponsor, defined, 260. *See also* stakeholders
Stake, Robert, 21, 121, 191–197
Stakeholder-based evaluation, 204
Stakeholders. *See also* bias; clients; conflict of interest; cultural competence; political considerations
  categories of, 317
  CIPP model, 176
  communication with, 65, 75–78
  data analysis and interpretation, 444, 447–449
  decision-oriented approaches, 185
  defined, 260
  definition of, 9
  describing, example, 303–304
  educating, 264
  evaluability assessment, 269
  expectations and understandings, 394–395
  human subjects, protections for, 86, 90–94
  identifying, 287–290
  interviewing, 298–299
  needs assessment, 316–321
  observations, data collection, 423–427
  participant-oriented approaches
    contemporary approaches, 199–204
    democratically oriented approaches, 215–219
    developmental evaluation, 207–208
    empowerment evaluation, 209–215
    naturalistic and fourth generation approaches, 197–199
    overview, 123, 189–191
    practical participatory evaluation, 205–207
    responsive approach, 191–197
    strengths and limitations, 223–226
    uses of, 220–222
  perceptions of, 300–301
  political environment, working in, 72
  program-oriented approaches, 167
  purpose of evaluation, 260–265
  question selection decisions, 330–332
  reports, 454, 458–461, 466–469, 485–486
  research on involvement, 220–222
  respect for, 87
  selection of, 200–204
  theory-based approaches, 161
  transformative paradigm, 116–117
  utilization-focused evaluation, 179–182
Standardized testing
  American Evaluation Association response to, 55
  data collection process, 439–442
  history of, 40, 42–43
  No Child Left Behind Act, 54–56, 165, 334, 440
  strengths and limitations, 166
Standard & Poor's rating agency, 78–79
Standards. *See also* Joint Committee on Standards for Educational Evaluation
  agreements and contracts, 367–368
  Countenance Framework, 191–192
  cultural competence, 232–233
  development by panels, 138
  evaluation criteria, sources for, 324
  evaluation profession, 47–49
  Guiding Principles for Evaluators, text of, 501–504
  objectives-oriented approaches, 154–159
  political validity, 71
  product evaluation, 146
  Program Evaluation Standards, text of, 499–501
  report conclusions and recommendations, 474–475
  specifying evaluation criteria and standards, 332–335
Standards-based approaches, 123, 165
Standards for Evaluations of Educational Programs, Projects, and Materials, 48
Standards of Evidence, 324, 370
Stanford Teacher Educational Program (STEP), 157–158
State government. *See* government programs
State licensing agencies, 127
Statewide testing, 40. *See also* No Child Left Behind (NCLB) Act (2001)
Statistical techniques
  data analysis and interpretation, 445, 446
  data collection designs, 343
  data evaluation, planning for, 354
  reports, presentation of, 473
  sampling, 407–411
3-step approach, 211
10-step Getting to Outcomes (GTO), 211
Storytelling, reports, 464, 477
Stratified random sampling, 409
Structured observations, 423, 425–427
Stufflebeam, Daniel, 173–177
Style, reports, 458–460, 464
Subcontractors, cost of, 366
Subject-matter experts. *See* expertise-oriented evaluations
Suchman, Edward, 160–164
Summaries, executive, 470–471
Summative evaluations
  accreditation, 134
  appropriate use of evaluation, 267
  bias and, 104
  CIPP model, 175
  data collection design and, 344
  evaluability assessments, 182–184

internal *vs.* external evaluator, 271
metaevaluation, 368–374
overview, 20–25
purpose of, 304
reports, 454–455, 461
stakeholder-based evaluations, 204
Supervisors, as evaluators, 51–53. *See also* stakeholders
Supplies, cost of, 365–366
Surveys. *See also* interviews
cross-sectional designs, 392–393
data collection, 406
developing, 428–429
electronic surveys, 432–433
face-to-face, 433–434
overview, 427–428, 442
telephone surveys and interviews, 431–432
Sustainability, empowerment approach, 211
Sweden, 50
Switzerland, 82
Symbolic use of findings, 480
Systematic inquiry, standards, 501
Systemization, 41, 87
Systems assessment, 179

Tables, reports, 473
Task Force on High Stakes Testing in K-12 Education, 55
Taxonomies
educational objectives, 42
essential competencies for program evaluators, 278–279
program-oriented approaches, 165
Teacher Education Accreditation Council (TEAC), 132, 136
Technology resources
data analysis and interpretation, 446
planning and, 307
reports, 460–461
surveys and, 430, 432–433
Telephone surveys, 431–432, 442
Tenure decisions, 127
Terminal objectives, Discrepancy Evaluation Model, 157
Testing
Buros Institute of Mental Measurements, 149
data collection process, 439–442, 442
program-oriented approaches, 165
Testing, norm-referenced, 40, 42–43
Theory, defined, 112n, 244
Theory-based approaches, evaluation, 123
heuristics of, 322
overview, 153–154
planning considerations, 309
program description, creation of, 292–298
program theory development, 298–300
strengths and limitations, 166–169
Theory failure, 161
Theory of valuing, 120
The Standards for Educational and Psychological Testing, 149
Thick descriptions, defined, 118, 344
Thorndike, E. L., 40
Time, influence on findings, 483
Timelines
data analysis, 354
data collection, 353
evaluation management, 358–366
evaluation planning, 261, 355–357
reporting, 454–455, 461–463
Time-series designs, 344, 388, 393–394
Time spent, evaluation tasks, 195
Title I, Elementary and Secondary Education Act (1965), 45
Total Design Method, 430
Total quality management (TQM), 57, 211
Training
data collection, 352, 439–442
empowerment evaluation, 210
evaluation personnel, 52, 305–307, 363–364
of evaluators, 492
logic models, 293
Transactions, Countenance Framework, 191–192
Transferability, participant-oriented approaches, 197
Transformative approaches
capacity building, 235–241
data collection methods, 350
empowerment evaluation, 209–215
paradigm of, 116–117
transformative participatory evaluation (T-PE), 201–202, 204, 322
Transparency, standards for, 86
Travel costs, 364–365
Trends, determining
cross-sectional designs, 392–393
data collection designs, 344, 401
time-series designs, 393–394
Trends, future of evaluation, 490–496
Triangulation, data collection, 385–386, 422–423
Truth tests, 481
Tuskegee Syphilis Study, 91
Tyler, Ralph W., 154–155
Tylerian evaluation approach, 154–155

UCLA evaluation model, 178–179
Unintended influences, report findings, 482
United Kingdom, 50, 131, 215
United Kingdom Evaluation Society, 49–51
United States, 50, 67, 130, 215. *See also* government programs
United Way of America (UWA)
internal evaluations, 51–53
logic model approaches, 160
outcomes measurement, 58–59
program theory and logic models, 292
use of evaluation, 29
University of Wisconsin Extension Program Development and Evaluation Office, 293
Unstructured observations, 423–424

Urban Institute, 46
U.S. Centers for Disease Control and Prevention, 237
U.S. Defense Department, 165
U.S. Department of Agriculture, 183
U.S. Department of Education
  accreditation, 133–136
  blue-ribbon panels, 138–139
  What Works Clearinghouse, 148
U.S. Department of Health, Education, and Welfare (HEW), 182, 268
U.S. Department of Health and Human Services, 90–91, 268
Uses of report information, 479–486
Utility
  defined, 11
  of evaluation, overview, 18–20
  role of evaluator, 16–17
  stages of evaluation theory, 120
  stages of theory evaluation, 122
  standards for, 84–85, 97, 499
  uses of report information, 479–486
  utility tests, 481
Utilization-Focused Evaluation (UFE)
  heuristics of, 322
  overview, 123, 179–182, 207–208
  strengths and limitations, 184–185
  use of results, 264
  uses of, 184

Validity
  consumer-oriented approaches, 144–145
  Discrepancy Evaluation Model, 157
  information sources and, 346–347
  mixed methods research, 386, 406, 422–423
  multicultural validity, 232
  participant-oriented approaches, 197–199, 223
  politics of evaluation, 70, 71
  questions selection, 317–318
Values
  bias and, 97–98
  constructivist paradigm, 116
  cultural competence, 232
  ethical standards, 96
  evaluation profession, 47–49
  stages of theory evaluation, 122
  theory of valuing, 120
  value-critical evaluators, 70
  value-neutral evaluators, 69–70
  value-sensitive evaluators, 70
Visual displays, reports, 465–466, 473, 478
Vouchers, schools, 56
Vulnerable populations, 91

Wandersman, Abraham, 209–215
War on Poverty, 43–44
Weiss, Carol, 121, 160–164
Welfare of public, standards for, 15, 87, 89–90, 503–504
Western Association of Schools and Colleges, 131n
Western Michigan University, Evaluation Center, 158–159, 177
What Works Clearinghouse (WWC), 148
Wholey, Joseph, 121, 182–184, 268
Wine tasting, connoisseurship, 141
W.K. Kellogg Foundation, 160
Working Group on Policy and Program Evaluation, 44
Work products, data collection, 441
World Bank, 237
Worth, judgments about, 47, 173–179
Writing style, reports, 464